

ARTIFICIAL NEURAL NETWORKS IN PUBLIC POLICY: TOWARDS AN ANALYTICAL
FRAMEWORK

by

Joshua A. Lee
A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Public Policy

Committee:

Laurie Schintler, Chair

David Hart

Michael Hunzeker

Vinodkumar Prabhakaran,
External Reader

Mark J. Rozell, Dean

Date: 4/24/20_____

Spring Semester 2020
George Mason University
Fairfax, VA

Artificial Neural Networks in Public Policy: Towards an Analytical Framework

A dissertation submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Public Policy

By

Joshua A. Lee
Masters of Arts
American University, 2013

Chair: Laurie Schintler, Professor
Schar School of Policy and Government

Spring Semester 2020
George Mason University
Fairfax, VA

Dedication

This dissertation is dedicated to my brother Zachary Lee and my father David Lee. May their memories continue to live on in those that knew them.

Acknowledgements

There are many individuals I need to acknowledge for this dissertation without whom it would not have been possible.

First and foremost my mother, whose constant support and encouragement have helped me more than she knows.

My Dissertation Committee members, including my Chair Dr. Laurie Schintler, as well as Dr. David Hart, and Dr. Michael Hunzeker. Their ideas, guidance, assistance, suggestions, and time have helped transform an assorted chunk of research pages into a finely-tuned piece of scholarship.

My External Reader, Dr. Vinodkumar Prabhakaran, who gave me fantastic insights into the field and consented to provide me tremendous help from nothing but a cold call.

My peer review/expert interviewees, whom I could provide no financial recompense, but whose input worked to thoroughly transform the principles in this dissertation with insights from fields far beyond public policy.

Andy, Nancy, and Lauren, who helped me practice and all gave me fantastic advice to sharpen and improve my defense presentation.

All the researchers and scholars, too many to name here, whom I have corresponded with over the years via email, called on the phone, and just generally helped me to understand what I was getting into and what I needed to do.

Finally, all my friends, both within my PhD program and without. Apologies to all the times I had to cancel, the times I locked myself in my room with no human contact, the times I didn't return your calls or texts quickly, and the times I was frazzled beyond belief that you had to deal with. But your putting up with me has made this process that much easier.

Table of Contents

	Page
List of Tables	x
List of Figures	xii
Abstract	
1 Introduction	1
1.1 Statement of the Problem	2
1.2 Purpose of the Study	3
1.3 Significance of the Study	3
1.4 Where to Focus: Machine Learning vs. Artificial Neural Networks	6
1.5 Research Question	10
1.5.1 Sub-Question 1: What are the key “research threads” to analyze, and how do these threads complement or interfere with one another when developing ANNs and other ML systems?	11
1.5.2 Sub-Question 2: What principles for developing ANNs and other ML systems in public agencies already exist within so-called “Ethical AI” frameworks?	12
1.5.3 Sub-Question 3: How should the information gathered from answering the above questions be evaluated, iteratively improved, and finally integrated into a cohesive analytical framework?	12
1.6 Scope Limitations	13
1.6.1 Domestic US Focus	13
1.6.2 Human Focus	14
1.6.3 Research Threads Focus	14
1.7 Importance of Definitions	15
1.8 Structure of the Study	15
2 Definitions and Taxonomies of Artificial Intelligence	18
2.1 Taxonomies of Artificial Intelligence	18
2.1.1 Reasoning Taxonomy	19
2.1.2 Domain Taxonomy	20

2.1.3 Training Data Taxonomy.....	21
2.1.4 Algorithm Taxonomy	22
2.1.5 Labelling Taxonomy.....	23
2.1.6 Task Taxonomy	24
2.2 Key Definitions.....	25
2.2.1 An Inductive vs. Deductive Approach to AI.....	26
2.2.2 What is Optimization?	30
2.2.3 Artificial Neural Networks: A Subset of Machine Learning.....	31
2.2.4 Unique Strengths and Weaknesses of ANNs.....	36
2.2.5 Conclusion	41
3 Background	42
3.1 Early History: Foundational Papers and Conferences.....	42
3.2 The First AI Winter.....	46
3.3 Backpropagation and the Thawing of Winter	47
3.4 The Second AI Winter.....	52
3.5 Taking Machine Learning by Storm: 2006-2013	53
3.6 Conclusion	54
4 Literature Review	55
4.1 Accuracy	56
4.1.1 True Positive, False Positive, False Negative, and True Negative	58
4.1.2 Recall, Precision, and F1 Score	59
4.1.3 Informedness and Markedness.....	62
4.1.4 Accuracy for ANNs: Alternative Measurements	64
4.2 Explainability	65
4.2.1 Foundational Works	66
4.2.2 What Makes a Good Explanation?	67
4.2.3 Taxonomy of Explainability Techniques	68
4.3 Fairness.....	69
4.3.1 Foundational Works	71
4.3.2 Conceptions of Fairness: Spaces, Beliefs, and Mechanisms	71
4.3.3 Conceptions of Fairness: Parity, Equality of Odds, and Calibration	78

4.3.4 Taxonomy of Fairness Techniques	81
4.4 Robustness	83
4.4.1 Foundational Works	86
4.4.2 Robustness Certification Standards	87
4.5 Privacy	91
4.5.1 Foundational Works: Differential Privacy	92
4.5.2 Other Privacy Techniques: Forming a Taxonomy.....	93
4.6 Democratic Legitimacy	95
4.6.1 Foundational Works	96
4.6.2 Algorithmic Governance.....	97
4.6.3 The Activities of Democratic Legitimacy	100
4.6.4 Synthesizing Democratic Legitimacy Activities	108
4.6.5 Key Democratic Legitimacy-Inducing Activities Defined.....	113
4.6.6 Separate Research Thread vs. Democratic Legitimacy Activity	115
4.7 Excluded Possible Threads	116
4.7.1 Ethical AI.....	116
4.7.2 Law & Regulation.....	117
4.7.3 Behavioral Psychology.....	117
5 Research Methodology	121
5.1 Introduction.....	121
5.2 Stage One: Testing of Competing Research Threads	123
5.2.1 Defining Optimization	124
5.3 Stage Two: Ethical AI Framework Meta-Analysis.....	126
5.3.1 Listing of AI Frameworks Covered.....	127
5.3.2 Four Existing Sources of Frameworks.....	130
5.3.3 Principle Assessment Criteria	131
5.4 Stage Three: Produce Draft Analytical Framework.....	135
5.5 Stage Four: Evaluate, and Improve Analytical Framework	135
5.5.1 Evaluating the Draft Analytical Framework	135
5.5.2 Key Characteristics to Evaluate	137
5.5.3 Evaluation Methods Explained	141

5.5.4 Expert Interview & Peer Review Procedures	143
5.6 Stage Five: ATI Study Comparison and Finalize Analytical Framework	144
5.6.1 Differentiating Comparative Analysis from Literature Review	145
6 Research Findings	147
6.1 Stage One Findings	147
6.1.1 Democratic Legitimacy’s Bilateral Relationships.....	147
6.1.2 Exclusion of Explainability	151
6.1.3 Other Bilateral Relationships.....	152
6.1.4 Summarizing Bilateral Relationships	158
6.1.5 Beyond Bilateral Relationships.....	159
6.1.6 Conclusions.....	159
6.2 Stage Two Findings.....	160
6.2.1 Summary of Extracted Principles	161
6.2.2 Discussion on “Arguable” Principles.....	166
6.2.3 Conclusion	168
6.3 Stage Three Findings	168
6.3.1 Initial Categories of Principles	168
6.4 Stage Four Findings	169
6.4.1 Key Critiques.....	170
6.4.2 Summary of Key Improvements from First to Second Draft of Analytical Framework.....	178
6.5 Stage Five Findings	179
6.5.1 Identifying Key Structural Differences.....	179
6.5.2 Limitations of the ATI Study	180
6.5.3 Differing Conceptions of the Key Research Threads.....	181
6.5.4 Accuracy	183
6.5.5 Fairness (non-Algorithmic & Algorithmic).....	183
6.5.6 Explainability.....	187
6.5.7 Robustness	188
6.5.8 Privacy	189
6.5.9 Democratic Legitimacy (Transparency).....	189

6.5.10 Democratic Legitimacy (Human Autonomy)	189
6.5.11 Democratic Legitimacy (SDPR)	190
6.5.12 Democratic Legitimacy (Accountability).....	191
6.5.13 Democratic Legitimacy (Deliberation).....	191
6.5.14 Democratic Legitimacy (Maintainability)	192
6.5.15 Democratic Legitimacy (Interpretability)	193
6.5.16 Additional Concepts	195
6.5.17 An Example for Applicability	198
6.5.18 Summary of Key Improvements from Second Draft to Final Draft	198
6.5.19 Conclusion	199
6.6 Final Analytical Framework	200
6.6.1 Public Agency Manager Principles	204
6.6.2 General Sociotechnical Principles	218
6.6.3 Human Interaction Principles	232
6.6.4 Optimization Principles	240
6.6.5 Vendor Principles.....	248
6.6.6 Conclusion	253
7 Conclusions	257
7.1 Final Thoughts	257
7.2. A New Case Study: Clearview AI and Facial Recognition	258
7.3 Impact on Public Agency Behavior	261
7.4 Addressing Literature Gaps	262
7.5 Relevance to the Future	263
Appendix A-1: Original Framework, Pre-Interviews	264
A-1.1 General Principles	264
A-1.2 Human Interaction Principles	268
A-1.3 Optimization Principles	271
Appendix A-2: Second Draft Framework, Post-Interviews/Expert Review/Revisions....	276
A-2.1 Public Agency Manager Principles	277
A-2.2 General Sociotechnical Principles	280
A-2.3 Human Interaction Principles	287

A-2.4 Optimization Principles	293
A-2.5 Vendor Principles	299
Appendix B: An Example Artificial Neural Network in Public Policy	303
B.1 Problem Formulation	303
B.2 Vendor Negotiations	305
B.3 Data Extraction and Acquisition	305
B.3.1 Example Input Data	305
B.3.1 Advantages of the Dataset	309
B.4 Data Pre-Processing	310
B.5 Modeling, Testing, and Validation	311
B.5.1 Results of the Model	312
B.5.2 Additional Validation: Testing Other Models	313
B.5.3 Conclusions from Testing and Validation	314
B.6 Deploy, Monitor, and Reassess	315
Appendix C: Interviewees	316
References	317
Biography	335

List of Tables

Table	Page
Table 1 - Disease Predictions vs. Reality	58
Table 2 - Fairness Mechanisms vs. Worldview	78
Table 3 - Democratic Legitimacy in Input vs. Output	101
Table 4 - Summary of Activities: Acceptance vs. Rejection	104
Table 5 - Linking Legitimacy Activities to the AIA	111
Table 6 - Definitions of Democratic Legitimacy Activities	114
Table 7 - List of AI Ethics Frameworks	128
Table 8 - Bilateral Relationships of Research Threads.....	158
Table 9 - Extracted AI Principles	161
Table 10 - Summary of Actions Taken 1	171
Table 11 - Summary of Actions Taken 2	172
Table 12 - Summary of Actions Taken 3	173
Table 13 - Summary of Actions Taken 4	174
Table 14 - Summary of Actions Taken 5	174
Table 15 - Summary of Actions Taken 6	176
Table 16 - Summary of Actions Taken 7	176

Table 17 - Summary of Actions Taken 8	178
Table 18 - Outcome Fairness Study Comparison	185
Table 19 - Explanatory Techniques Study Comparison	188
Table 20 - Answerability vs Auditability.....	211
Table 21 - Confusion Matrix.....	312

List of Figures

Figure	Page
Figure 1 - Dimensions.ai analysis of scholarly publications.....	7
Figure 2 – Reasoning Taxonomy	19
Figure 3 - Domain Taxonomy	20
Figure 4 - Training Data Taxonomy	21
Figure 5 - Algorithm Taxonomy	22
Figure 6 - Labelling Taxonomy	23
Figure 7 - Task Taxonomy	24
Figure 8 - The Atari Pong game.....	28
Figure 9 - Basic Artificial Neural Network.....	32
Figure 10 - Changes in F1 Score by Recall/Precisions.....	61
Figure 11 - Adversarial perturbation of image	85
Figure 12 - Robustness Certification	88
Figure 13 - Model Factsheet Example	213
Figure 14 – DOHA Model Factsheet.....	215
Figure 15 - Correlation != Causation Examples.....	220
Figure 16 - Correlations != Causations (Old Example)	282

Figure 17 - TensorFlow Model Summary.....	312
Figure 18 - Comparison to Other ML Models	314

Abstract

ARTIFICIAL NEURAL NETWORKS IN PUBLIC POLICY: TOWARDS AN ANALYTICAL FRAMEWORK

Joshua Lee, Ph.D.

George Mason University, 2020

Committee Chair: Dr. Laurie Schintler

This dissertation assesses how artificial neural networks (ANNs) and other machine learning systems should be devised, built, and implemented in US governmental organizations (i.e. public agencies). While it primarily focuses on ANNs given their current prevalence and accuracy, many of its conclusions are broadly applicable to other kinds of machine learning as well.

It develops an *analytical framework*, drawn from diverse fields including law, behavioral psychology, public policy, and computer science, that public agency managers and analysts can utilize. The framework yields a series of principles based on my research methodology that I argue are the most relevant to public agencies. The qualitative methodology consists of an iterative approach based on archival research, peer review, expert interviews, and comparative analysis.

Critically, this dissertation's intent is not to provide the specific *answers* to all questions related to machine learning in public agencies. Given the speed at which this field changes, attempting to provide universally applicable answers would be difficult and short term at best. Rather, this framework focuses on principles which can help guide the user to the proper questions they need to ask for their particular use case. In that same vein, the normative principles it provides are procedurally focused in scope rather than focused on policy outcomes. In other words, this framework is meant to be equally applicable regardless of what one's specific policy goals are.

1 Introduction

Over the past decade, artificial intelligence (AI) has made incredible strides forward. Artificial neural networks (ANNs), a specific type of AI, can often perform a wide range of tasks that require human intelligence more accurately than any other kind of AI system preceding them. However, although ANNs and other machine learning (ML) systems have unique capabilities that previous AI systems cannot easily match, they also have unique limitations that must be simultaneously considered. These weaknesses are particularly relevant when considering issues of public policy and public administration, an intersection that has often gone ignored.

We stand at the precipice of a new kind of government for a wide array of public services. Utilizing ANNs has the potential for significant improvements to these services while also opening the door to new problems. In this study, I intend to address what these new problems are, how they could be mitigated, and how we should assess if they are being mitigated. I argue that what public agency managers and analysts need most is a framework of principles and questions for properly developing ANNs and other “black box” machine learning systems for use within their public agencies. By black box, I mean a machine learning system where interpreting the system’s inner workings is extremely difficult even for expert computer scientists, and where even experts cannot

achieve full explainability for a system's decisions. While this dissertation will be focused primarily at public agency managers and analysts, many of the points raised should be relevant to a broad array of social scientists and the private sector.

1.1 Statement of the Problem

There is already an extraordinarily large body of research concerning machine learning generally, and research into ANNs specifically has exploded since 2012 even compared to other advanced machine learning techniques such as support vector machines (SVMs) (Jeeva 2018). Research into ANNs has traditionally approached the subject from one of three levels of analysis:

Micro-level Analysis - Analysis that focuses on the technical specifics of artificial neural network architecture. This includes almost any research directly focused on maximizing the predictive accuracy of an ANN through improving the structure, data, algorithm, or training process of the ANN. This level of research is almost entirely from within the field of computer science.

Macro-level Analysis - Analysis that focuses on the impact of ANNs more broadly in society. This includes looking at the impact of autonomous vehicles, AI in the military or the judicial system, etc. This level is where much of the social science research related to ANNs resides.

Mezzo-level Analysis - Analysis that focuses on the broader patterns gleaned from innovations at the micro-level while also analyzing how they impact broader issues

faced at the macro-level. Examples include dealing with issues such as ethics, fairness, privacy, explainability, and robustness, among other issues.

1.2 Purpose of the Study

The purpose of this study, then, is to develop a mezzo-level *analytical framework* for public agency managers and analysts who work with and manage any kind of black box machine learning system, particularly ANNs. By analytical framework, I mean a series of normative principles and associated follow-on questions that I argue should be considered during the development and implementation of these systems for use in a governmental decision-making capability. Of note, these principles will be procedurally-focused rather than focused on policy impact and outcome: rather than attempting to assert that a given policy outcome is favorable, the analytical framework is focused on ensuring that key pitfalls are avoided and that the most important questions are being asked.

1.3 Significance of the Study

Artificial neural networks have already made waves in fields far beyond computer science, including many with public policy relevance. This includes political science (Weber, et al. 2017), healthcare (Raghupathi and Raghupathi 2017), law (Kehl, Guo and Kessler 2017), transportation (Bojarski, et al. 2016), and defense (Barker 2016), among others. However, while many fields are beginning to experiment with developing

and implementing ANNs, there is little in the way of coherent guidance outside of computer science. Nowhere is this gap clearer than in the field of public policy.

Unfortunately, computer science micro-level analysis is often technical to the point where those who are not specialists themselves cannot even understand the gist of the material. Indeed, many computer scientists tend to abstract problems into almost purely mathematical terms, which may not be the ideal mechanism to convey their ideas to a wider audience in the social sciences (Selbst, et al. 2019). Even in much of the mezzo-level literature, a reader will often need to be well versed in mathematical notation and a wide array of niche terminology that most readers will lack. Simply put, there is a lot of existing computer science literature is unapproachable for too many outside the field.

There are already plentiful examples of public agencies essentially implementing ML systems in whatever way they so choose, treating them indistinguishably from generic software systems and vendor contracts that public agencies have dealt with for decades. Indeed, we don't even know precisely how far many US public agencies, including those at the state and local level, are going in their use of these systems due to a fundamental lack of transparency (Brauneis and Goodman 2018).

In fairness to those public agencies, however, there is little evidence to indicate that most of this activity is done with inherently malicious motives or that these agencies are attempting to achieve goals entirely outside of their agency's scope. We

can see this readily with the FBI's usage of ML systems in DNA fingerprinting (Abadicio 2019) or facial recognition technology (Government Accountability Office 2016). On the one hand, few would argue that the FBI is stepping out of its legitimate authorities through matching DNA evidence or faces at crime scenes to databases of criminals. On the other hand, the usage of ML systems to attempt to achieve these legitimate goals can lead to a wide range of problems that are not easily visible without deep and considered study.

In short, what we all too often have today (not just with the FBI, but for many public agencies at the local, state, and federal level) is generally good intentions combined with carelessness, lack of forethought, and no clear guidance from those who oversee these agencies as to what should be permissible and how they should go about using ML systems. Unfortunately, when dealing with ANNs and other black box ML systems, these issues combined can equate to just as much damage to the public as malicious intent.

Of course, all blame should not be laid at the feet of public agencies alone: even if a public agency's leadership knew about and wanted to mitigate these issues, there is scant guidance for public agency managers to follow. Almost all the computer science literature is entirely unapproachable, and most macro-level scholarship focuses on the broad abstractions of societal impact rather than the nuts-and-bolts of implementation. All of this, then, leads to the significance of this study: to generate a framework which

provides such guidance to public agency managers and analysts so that they can avoid such pitfalls when developing these systems.

1.4 Where to Focus: Machine Learning vs. Artificial Neural Networks

One constant question that must be dealt with early (and indeed has already shown itself in the sections above) is what kinds of algorithms should be focused on: machine learning algorithms generally, or artificial neural networks specifically. Some literature in the field focuses explicitly on artificial neural networks, some on machine learning generally, and some even on algorithmic decision-making more broadly still to include any kind of automated decision-making. Additionally, while almost all scholarship agrees that artificial neural networks are a subset of machine learning, there is little agreement on the bounds of machine learning itself; both scholarly and non-scholarly writing may refer to a wide variety of very different methods under the umbrella of “machine learning”.

To better grasp why I have a particular focus on ANNs, first consider the trends in scholarly AI publications over the past decade:

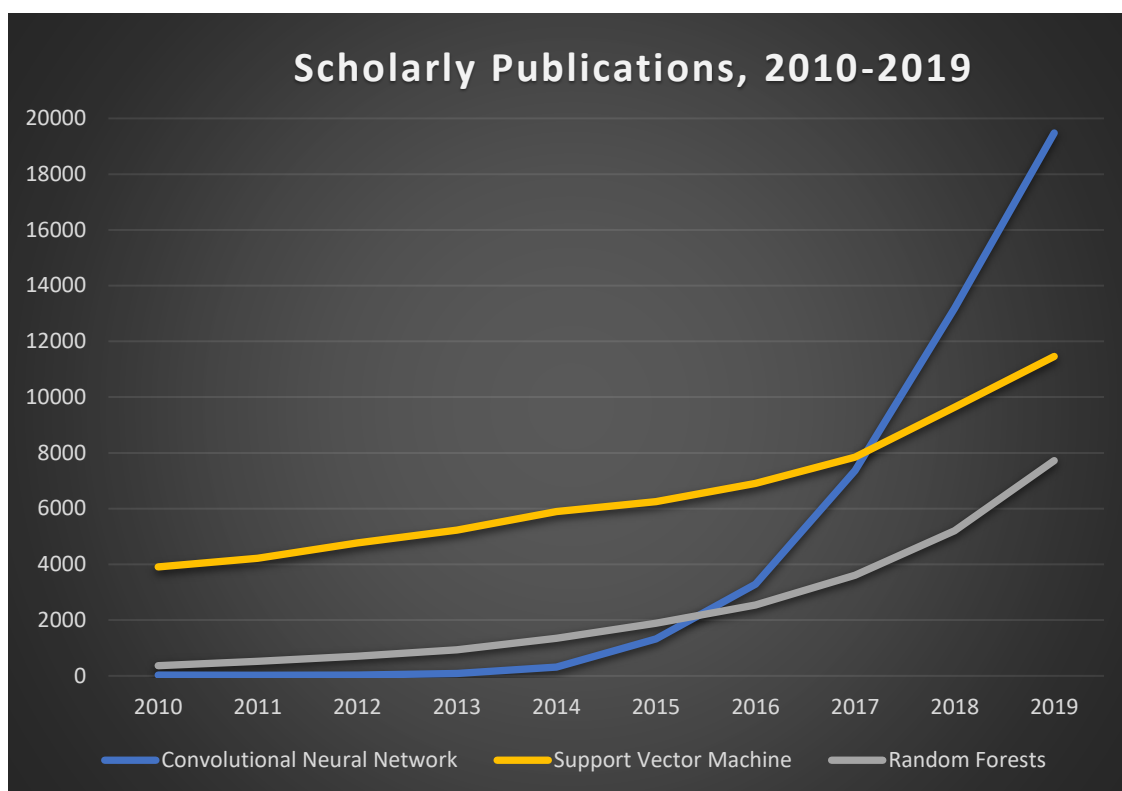


Figure 1 - Dimensions.ai analysis of scholarly publications¹

As we can see above in *Figure 1*, neural network research has exploded since 2016 and shows no signs yet of slowing down (as of January 1st, 2020, at least). If anything, *Figure 1* likely *understates* the amount of artificial neural network literature since it only includes convolutional neural networks. In contrast, literature into support vector machines and random forests (two other popular advanced machine learning techniques) are also growing, but significantly more slowly.

¹ Query for Convolutional Neural Networks : "convolutional neural network"; Query for Support Vector Machines: "support vector machine"; Query for Random Forests: "random forests"; Queries done on "Title and Abstract", not Full Text

On the one hand, choosing artificial neural networks alone as the focus may unnecessarily minimize this study's potential relevance: many of the principles and ideas that apply to ANNs are also applicable (albeit to varying degrees) to other kinds of machine learning. Thus, to ignore all other machine learning literature simply because the literature was not *explicitly* focused on artificial neural networks would potentially exclude countless relevant research articles, not to mention highly relevant background history.

On the other hand, attempting to include the entirety of machine learning, broadly defined, is also inherently problematic. For one thing, there is no absolute accepted definition as to what machine learning encompasses. For example, do mathematical algorithms initially developed in early the 19th century, such as linear or logistic regression, qualify as machine learning? While logistic regression can be a form of a basic one-layer neural network (Raschka 2019), linear regression is often excluded from the category entirely. Even ignoring those cases, there are more than a dozen relevant, broad sub-categories within machine learning (artificial neural networks being just one of them), most with further variants and sub-variants (S. Ray 2017). While they all may share some pitfalls, they also all have traits which may be unique to varying degrees. The potential scope of attempting to include the entirety of machine learning in such a study alongside ANNs is daunting, to say the least.

This study tries to find a balance between the two poles: to draw on earlier research focused on other machine learning techniques for background and foundational literature while *primarily* focusing on artificial neural networks. When this study *does* focus on machine learning, except where explicitly noted, it should be understood as a relatively narrower definition of machine learning techniques to include only those that are generally ‘black box’ in nature; examples of such black box machine learning systems include support vector machines (SVMs), random forests, and of course ANNs.

What makes these ML systems black boxes is that it is not clear how they produce the results that they do; even though we can spell out the math that goes into each of them, that doesn’t mean we understand *why* a particularly complex combination of mathematics just so happens to often produce the desired answer most of the time. To make matters even murkier, in several cases this study discusses the public agency in question did not reveal precisely which machine learning algorithm it used; such cases create an *administrative* black box that is just as impenetrable (if not more so) than the algorithmic black boxes of these techniques, and will be considered black box machine learning techniques for the purposes of this study.

1.5 Research Question

In this dissertation, I ask a primary research question (immediately below), as well as three sub-questions which will help to answer the primary question. The primary research question is as follows:

What questions and principles should guide public agency managers when developing artificial neural networks?

There are several word choices in this research question that bear further examination. First, the term “developing” is important and meaningful. There are a wide variety of different verbs which could be used here, with most of the tension arising from the line between *construction* of an ANN and *implementation* of an ANN. That is, simply building the ANN versus applying the ANN. I concluded that the most relevant term for use is from the realm of computer science. The term “developing”, from software development, is used because the software development lifecycle (SDLC) is relevant both because an ANN is inherently software, and because the SDLC itself is similar to many public policy development processes (Stackify 2017).

Another term worth noting is “manager.” Determining *who*, precisely, this framework is meant for is important: policy scholars in academia, or public policy administrators in government. In general, the target audience should be the managers in public agencies who will be responsible for developing the proposed ML system, regardless of whether it is developed in-house or through external vendors.

Additionally, public agency analysts are also targets as those individuals most likely to be using these systems internally on a daily basis.

Finally, the phrase “questions and principles” is undeniably vital. The result of this study will be the development of a series of both principles and questions which will help guide public policy managers and analysts conceive of, design, implement, and maintain ML systems. It is equally important to note that the words “answer” and “solution” are not present in my research question - the field of ANNs is evolving so quickly and in so many directions that even if I were to arrive at “correct” answers today, they could easily be wrong within a few months. In contrast, determining the proper questions that need to be asked during development should remain more constant and useful over time.

1.5.1 Sub-Question 1: What are the key “research threads” to analyze, and how do these threads complement or interfere with one another when developing ANNs and other ML systems?

The term “research thread” is one that is used frequently in this study. It denotes a distinct area of machine learning scholarly research which may intersect with other research threads but is nevertheless distinct in its scholarly origins and the key elements it focuses on as important. The first step of this study, then, is to identify the most important research threads related to mezzo-level analysis of ANNs and other ML systems.

Beyond a simple recitation of a literature review, however, this study focuses more explicitly on the interaction *between* these research threads. For example, does improving (or *optimizing for*, as the case may be) one of the research threads cause damage to the desired end state of a different research thread? Alternatively, do they play a complementary role to one another wherein improvement to one thread is improvement to another?

1.5.2 Sub-Question 2: What principles for developing ANNs and other ML systems in public agencies already exist within so-called “Ethical AI” frameworks?

This question revolves around archival research and comparative analysis: in the past three years, dozens of “ethical AI” frameworks have been released by a whole range of entities and groups including those in the commercial sector, individual scholars, scholarly institutions and conferences, and even governments. While almost all of this scholarship is not aimed at public agencies specifically, many of the frameworks nevertheless discuss key issues of ML development in different circumstances.

1.5.3 Sub-Question 3: How should the information gathered from answering the above questions be evaluated, iteratively improved, and finally integrated into a cohesive analytical framework?

With the above questions answered, the last step is to bring it together into a cohesive analytical framework. Simply making a bullet point list of “good ideas” from each of the previous sub-questions, or even from each of the sections of the literature

review, is insufficient. Rather, there needs to be evaluative mechanisms to determine whether a normative principle is worthy of inclusion.

1.6 Scope Limitations

This study has several limitations on its scope and applicability. Some of the scope limitations were chosen to ensure thoroughness would not be lost for the sake of covering excess topics, while others were necessary to spell out where and when this study should be applicable.

1.6.1 Domestic US Focus

This study will concentrate on domestically focused public agencies within the United States. I chose this scope for several reasons. First, this is a procedurally normative study based on the rules and norms governing US public agencies. Therefore, its conclusions may not apply equally to public agencies in different countries. Different cultures and different political systems can have not only different societal values, but also different ways in which public agencies function. Indeed, what is “fair” in one country may not be deemed fair in another due to differing cultural norms and/or history.

Second, this study focuses on domestically focused public agencies because it touches on an individual’s rights. US citizens (and to an extent anyone located within the United States) are entitled to different protections under US law than those located outside the United States. As such, US public agency activities aimed outside the United

States (from entities such as the US intelligence community, the Department of Defense, the Department of State, etc.) fall outside the scope of focus for this framework.

1.6.2 Human Focus

This study will focus on ML systems which assess human beings and their actions and/or decisions. This includes both systems with input data related to people in society – social, economic, political, religious, etc. – and systems which require human beings to frequently use them and be assessed by them in some manner. I chose this focus because public policy itself is concerned predominantly with people. In contrast, dealing with machine learning techniques aimed at predicting a tree’s height from its width are perhaps fascinating, but much less relevant to domestic public policy analysis.

1.6.3 Research Threads Focus

Simply put, there are too many mezzo-level research threads to reasonably focus on all of them simultaneously in my study. While all of them are important, some have particularly high relevance for public policy analysis, and others are located outside of public policy with little in the way of established scholarly research. Because of this, this study will primarily focus on six key research threads: accuracy, fairness, explainability, robustness, privacy, and democratic legitimacy. The decisions for why these threads were chosen (and others weren’t) for the literature review and further analysis will be discussed in the literature review section below.

1.7 Importance of Definitions

Even while attempting to minimize the amount of mathematics necessary to understand this study, there are nevertheless several critical concepts that must be understood, if only at the abstract level. These concepts include models, layers, neurons, activation functions, and training and testing data, among others. The next chapter will go into these in greater depth.

1.8 Structure of the Study

This study is structured into seven chapters and three appendices, with this first chapter as the current Introduction. The second chapter (Definitions and Taxonomies of Artificial Intelligence) provides the most important terminology definitions for understanding this study and identifies the various taxonomies for artificial intelligence. It explains the key mathematical components that make up an ANN, as well as where they fit into the various taxonomies of AI. Understanding the terminology and different taxonomies available will also help in one's understanding of the literature review ahead, since different literature can use different taxonomies as its lens through which to analyze a given ML system.

The third chapter (Background) dives into the history of AI and ML systems, particularly ANNs, and specifically as that history relates to public policy and the federal government's involvement. It provides a historical grounding and an introduction for the literature review as well.

The fourth chapter (Literature Review) dives into each of the six key research threads I noted above. It focuses on their foundational works, scholarly debates, taxonomies of concepts within the field, and finally how that thread is normatively assessed (i.e. more vs. less accurate, more vs. less fair, etc.). It also has a section on the research threads that were *excluded* from the literature review and why.

The fifth chapter (Research Methodology) dives into how I intend to conduct my study and why I use the methods I do. The research methodology section is split into five Stages (discussed further below), each corresponding to one or more research methods being used in the methodology.

The sixth chapter (Research Findings) displays the results from following each stage of the research methodology. The last section in Chapter Six contains the actual analytical framework.

Finally, the seventh chapter (Conclusion) looks at the broader implications of my Research Findings and their applicability to the future. In addition, it looks at ongoing discussions in society revolving around ANNs and what this analytical framework might say about them. It concludes with a brief look at the many future areas of potential research that should follow up this study.

The three appendices afterward are meant to provide additional illumination for those interested in digging deeper into the research. Appendix A-1 shows the state of the initial draft analytical framework after the completion of Stage Three, and Appendix

A-2 shows the state of the second draft analytical framework after the completion of Stage Four. Appendix B provides an “example” artificial neural network created by me (which I refer to as the “DOHA model”) specifically to help explain and apply the analytical framework in Chapter Six. Finally, Appendix C notes the interviewees who assisted with Stage Four of the analytical framework. Of note, only the final analytical framework (after Stage Five) is shown in the Research Findings chapter.

2 Definitions and Taxonomies of Artificial Intelligence

In this section, I provide two key elements necessary for understanding my analytical framework (not to mention the background and literature review sections). First, this section provides the major taxonomies for defining artificial intelligence at the broadest level – how AI is defined in different ways, what those definitions mean, and how they differ from one another. Second, this section includes more in-depth definitions and concepts that are necessary to understand ANNs and ML systems generally at an abstract level. While the taxonomies section touches on definitions briefly, the latter section goes into greater depth.

2.1 Taxonomies of Artificial Intelligence

While each taxonomy below is not mutually exclusive, they nevertheless each work to differentiate AI systems based on a different set of characteristics. Of note, I also provide literature on other taxonomies throughout the literature review. I place this continual emphasis on taxonomies for several reasons. First, this study is at the intersection of at least two fields: computer science and public policy. While much of its content discusses complex ideas from computer science, its primary audience is meant for those in public policy. Because of this, I argue that the best way to understand these complex concepts without the requisite computer science background is to

conceptualize these research threads (as well as AI itself) into various taxonomies. In this way, a complex concept can be broken down and the differences between elements within the concept can be better understood without needing linear algebra.

Additionally, this field is replete with taxonomies already; no new taxonomies were needed to be invented for this dissertation. Indeed, just about every concept discussed in this study already has its own taxonomy of ideas (and at times several competing taxonomies). If anything, the difficulty was in choosing which competing taxonomy is the most useful. This allows me to keep my primary focus on my research questions rather than defining and then justifying self-created taxonomies. While this chapter sometimes provides new names to the taxonomies it shows, the actual taxonomies themselves are well documented in other research.

2.1.1 Reasoning Taxonomy

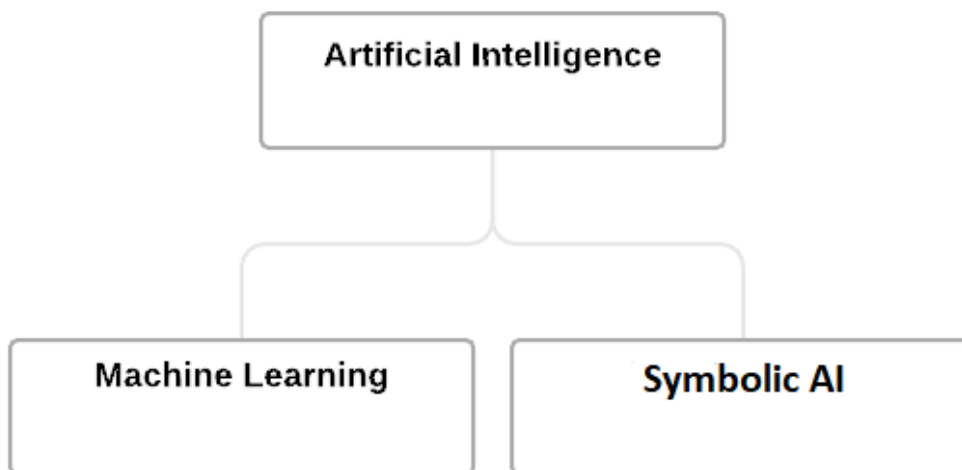


Figure 2 – Reasoning Taxonomy

The *Reasoning Taxonomy* splits the field of AI into two halves: machine learning and symbolic AI. As we can see from *Figure 2* above, both machine learning and symbolic AI are subsets of artificial intelligence. Expert systems is the most well-known type of symbolic AI technique, and artificial neural networks are a type of machine learning technique. This study almost exclusively focuses on machine learning rather than symbolic AI, except for background information in Chapter Three.

The basic distinction between them is based on inductive vs. deductive reasoning: whereas an expert system/symbolic AI is provided rules to follow and then applies those rules to a given data set, a machine learning system learns the rules based on being shown individual examples first. In this way, they are opposites of one another – expert systems are made directly with human-designed functions, whereas machine learning systems develop their own functions based on examples.

2.1.2 Domain Taxonomy

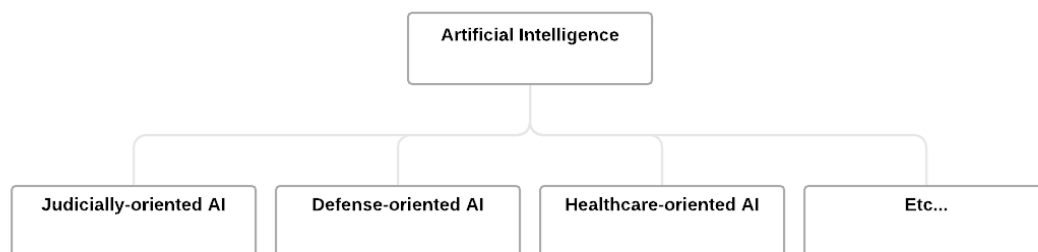


Figure 3 - Domain Taxonomy

The *Domain Taxonomy* should be the most intuitive: it simply splits AI systems by what subject matter they are designed to focus on. The AIs themselves could be based

on symbolic AI or machine learning, but in this taxonomy they are split solely based on what subject matter they focus on in terms of input and output data. For example, a judicially-oriented AI system would be like COMPAS, which is a machine learning tool used in the Wisconsin judicial system to determine who is a likely risk of recidivism among prison inmates (Angwin, et al. 2016); this particular example will be discussed throughout the dissertation at different points.

2.1.3 Training Data Taxonomy

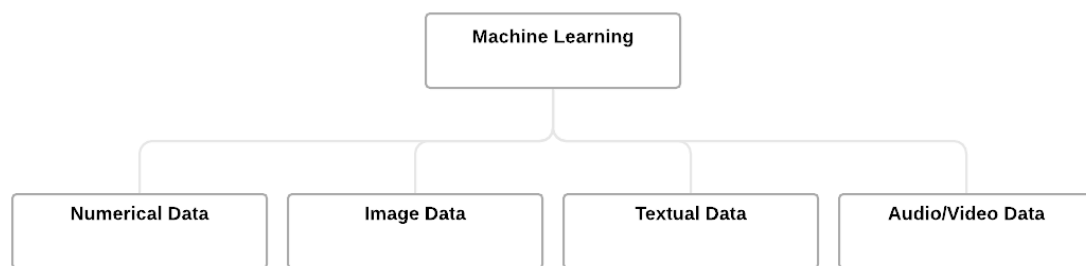


Figure 4 - Training Data Taxonomy

This taxonomy only applies to machine learning, not expert systems/symbolic AI – symbolic AI does not *have* a data taxonomy since they are not “trained” on data. The training data taxonomy focuses on the type of data fed into a machine learning system.² Numerical data would include anything inherently numerical in nature, such as basic

² While all inputted data to an ANN ends up being converted into numerical data of some type, in this case I am referring to data that is *originally* numerical in nature – percentage values, categories, scales, etc.

true/false (i.e. Boolean) information, quantities and amounts, and percentages. Image data would include data such as the locations and colorings of pixels in an image. Textual data would encompass just that, but generally encoded into some form of numerical representation of the text's meaning and/or position; audio and visual data would be similarly encoded.

2.1.4 Algorithm Taxonomy

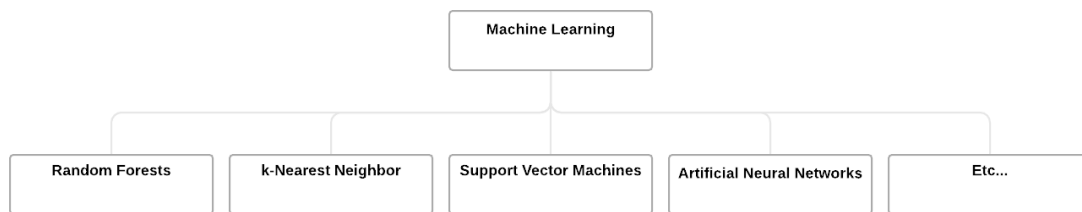


Figure 5 - Algorithm Taxonomy

The *Algorithm Taxonomy* is also relevant only to machine learning. It splits machine learning techniques into their actual algorithms. There are many more machine learning algorithms than are listed above, not to mention sub-variants for each kind. This taxonomy is differentiated by the nature of the algorithm: what kind of math is performed on the input to achieve the output, and how training data impacts changes in the math. While this dissertation has a focus on artificial neural networks, it also looks at other common black box machine learning algorithms in the algorithm taxonomy.

2.1.5 Labelling Taxonomy

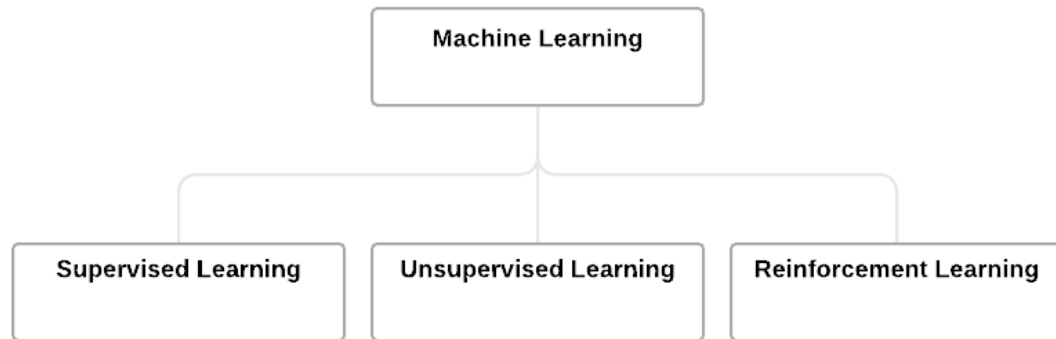


Figure 6 - Labelling Taxonomy

The *Labelling Taxonomy* focuses on how the algorithm uses labels in the data.³

The key difference between the different methods is how the data they are trained on is labelled. Supervised learning techniques generally accept data that is entirely pre-labelled in advance – for example, a dataset of images where each image is clearly labelled (by a human) as to whether it has a cat or a dog in it. In contrast, unsupervised learning (sometimes also referred to as self-supervised learning) would be when you have the same dataset of images but you *don't* know in advance whether or not the image has a cat or a dog in it. Rather, the unsupervised algorithm might attempt to cluster images it deems similar into different groups. The types of algorithms (from the algorithm taxonomy) that can be applied to supervised learning vs. unsupervised

³ There are also other less common methods in the labelling taxonomy besides these three, such as one-hot learning and semi-supervised learning

learning techniques are generally different, although artificial neural networks have both variants. Presently, supervised learning is more common for finding end results, whereas unsupervised learning is more common for various kinds of data pre-processing.

Finally, reinforcement learning techniques fall somewhere in between – the label is generally only known after a given decision has been made, and that delayed feedback is then constantly used to have the algorithm learn. One of the most common examples of reinforcement learning being used in the real world actually involves video games. In these cases, the “label” would be how good or bad a given action is in the video game, which often is not known until seconds, minutes, or even hours have passed. One of the most recent and high-profile examples was with the video game Dota 2, where five AI bots trained with ANN-based reinforcement learning played as a team to defeat a team of five top professional human beings (Statt 2019).

2.1.6 Task Taxonomy

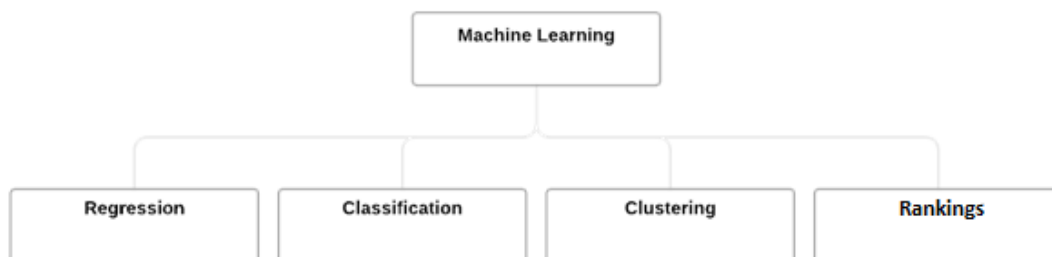


Figure 7 - Task Taxonomy

Finally, the *Task Taxonomy* splits machine learning systems by what kind of output they're producing. A regression output is predicting a continuous variable (such as one's height) whereas a classification output is predicting a discrete variable (such as True vs. False). A clustering output is the partitioning of the input variables into multiple groups. Finally, a ranking output orders different inputs for importance; recommender systems, such as when Netflix suggests a movie for you to watch, are a subset of rankings. This taxonomy is also closely tied to the labelling taxonomy – regression and classification are generally tied to supervised machine learning, clustering is generally tied to unsupervised machine learning, and ranking systems can be tied to any of the three.

2.2 Key Definitions

One key issue for this study is how deeply to dive into the technical minutia of machine learning and artificial neural networks. On the one hand, this is not a computer science dissertation, and most of the readers are not expected to be computer science scholars. At the same time, there are terms that are inherently mathematical in nature that require at least an abstract grasp of in order to understand the content of this dissertation. My intention when constructing this study is to limit the needed technical definitions (save for some elements of the literature review) such that those with a grasp of algebra and statistics will be able to follow. While some of the information below may be repetitive from the taxonomies section above, this is done to provide multiple avenues for understanding the concepts.

There are a multitude of ways to define artificial intelligence. In the field of computer science, it is generically defined as *a device or program that perceives an environment and acts upon that environment to achieve some goal*. For practical purposes, all this means is a computer program which accepts a certain input (such as numbers, a picture, text, etc.), and based on that input produces an output that the user is interested in, such as classification of the input into categories or a probability calculation.

2.2.1 An Inductive vs. Deductive Approach to AI

Symbolic AI techniques such as expert systems use a *deductive approach* to building an AI. That is, based on a set of general principles coded explicitly into the AI by a human expert, the AI then reacts to the provided input. This is also known as a top-down approach – the AI can only ever be as good as the human subject matter experts that designed it. One excellent example of the symbolic AI approach is Deep Blue, the AI which defeated chess legend Garry Kasparov (Greenemeier 2017). Deep Blue had human chess experts and computer scientists team up to write gameplay rules and utilize the brute computing power available to scan deeply ahead for possible moves to make.

In contrast, machine learning techniques use an *inductive approach* to building an AI. They consist of computer algorithms which utilize a dataset to “learn” about a

given problem, then use what they've learned to make future predictions. This is the opposite of the symbolic AI approach in many respects – rather than a human expert determining how the AI should function, machine learning techniques form general principles based on a specific set of data provided to them in advance. This should not be confused simply with using a database, like Deep Blue did. Rather, what separates machine learning systems is that the algorithms they use to make predictions are themselves modified and (ideally) self-improving from running through these datasets to learn from.

Visualizing the Difference: Playing Pong

An easy way to visually conceptualize the difference between them is with an imaginary AI that plays the classic computer game *Pong*:

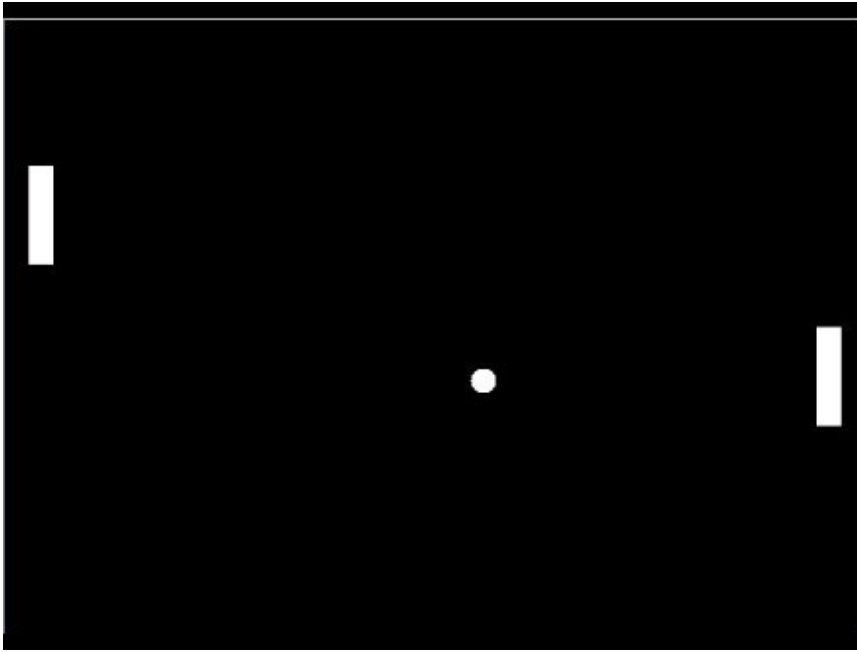


Figure 8 - The Atari Pong game

Pong is a simple and straightforward game – there are two players, and each player controls a paddle. The objective of the game is to hit the ball over to the opposing players side with your paddle without them being able to get their paddle there first to bounce it back. Paddles can only move up and down, and while the ball's angle will change depending on how it is hit on the paddle, its speed remains the same.

Both machine learning and expert systems techniques could be utilized to develop a “Pong AI”. An expert system Pong AI would have general principles (i.e. rules) coded into it. For example, a computer programmer could implement a simple rule for the AI that says “calculate the angles at which the ball will hit the top and bottom of the screen in order to place the center of your paddle where the ball will be when it reaches

your side.” This “rule” would likely enable an expert system Pong AI to play quite well.

We can see the deductive reasoning playing out - the system was simply told what to do by the human subject matter expert. Of course, for more complex situations these instructions can be vastly more complex and based upon complex statistical analysis as we saw with the chess AI Deep Blue.

In contrast, a machine learning Pong AI would have no such human-generated principles to guide it. Rather, it would be provided countless thousands (or millions) of example games to learn from. From those games, it would slowly learn which moves of the paddle were likeliest to increase its own score and least likely to increase its opponent’s score.⁴ Were you to observe the machine learning system while it *trained* on the data, you would likely laugh at its poor initial and seemingly random performance. Eventually, assuming the machine learning algorithm was well-designed, it would learn how to play pong with a high degree of proficiency.

For a relatively simple example like Pong, an expert systems approach would likely perform just as well as a machine learning approach, if not better. However, what if we were dealing with something far more complex than a simple decision to move the paddle up or down? For more complex tasks, human knowledge is unlikely to be perfectly explained with rules.

⁴ There are variants of machine learning which work somewhat differently than this, but the general concept remains the same.

2.2.2 What is Optimization?

Arguably the most important mathematical concept that needs to be understood to grasp what it is machine learning actually *does* is to understand the basics of optimization. All machine learning algorithms have an optimization function of one kind or another: a mathematical algorithm which attempts to find the greatest or least value(s) for an equation given some constraint(s). Below is a simple example of an optimization function:

Solve for the smallest value for A (i.e. $\min(A)$) for the following equation:

$$\min(A): 2B + 2C$$

Without any kind of constraint, of course, this would hardly be an optimization problem at all: the minimum value for A would be infinitely low, you'd just need to keep decreasing B and C. However, now let's add one more constraint to our function:

$$B * C = 1000$$

The constraint states that B multiplied by C *must* equal precisely 1,000. If this constraint is violated, any potential solution wouldn't count. Now, if you were to further restrict this to only analyzing whole number values for B and C, you could actually solve this through trial and error with basic math. With that additional stipulation, the equation would be optimized when B = 40 and C = 25 (or vice versa). In such a case, A = 130. However, were you to *not* restrict yourself to whole numbers, through calculus (which is beyond the scope of this dissertation) you would discover that the truly

optimized solution to the problem is when $B = \sqrt{1000}$ and $C = \sqrt{1000}$. In such a case, A would be equal to ~ 126.5 , a more optimized solution for our equation since A is smaller.

This is what is at the core of how almost all machine learning functions. The only difference, particularly exemplified with ANNs, is that there are millions or even billions of such parameters that must be optimized instead of just B and C. This perhaps helps to explain the “black box” nature of many machine learning systems: even for the comparatively simple optimization problem above, figuring out *why* the optimal solution occurs when B and C are equal isn’t immediately intuitive without a background in math. Now consider how difficult it would be to figure out the *why* given an equation with millions of such values, particularly if the algorithm used to try and optimize the values involved highly complex mathematics itself. In such a case, simply *seeing* all the values is insufficient to understand *why* together they would all produce the most optimal result. This is the essence of the black box problem in machine learning.

2.2.3 Artificial Neural Networks: A Subset of Machine Learning

While there are many explicitly *mathematical* definitions available for ANNs already (XenonStack 2017), as I stated earlier I intend to provide as math-free a definition as possible while still allowing for a conceptual understanding of how they function. Much technical detail is necessarily omitted from the definitions below for the sake of brevity and clarity.

Visualizing an ANN: Layers, Neurons, and Weights

At its core, an ANN is just another kind of machine learning technique, and a distant cousin of logistic regression. It attempts to very loosely simulate the human brain's neurons inside a computer – there are *layers* of connected artificial *neurons*, with each connection having a *weight*.

Consider the following visualization of an ANN:

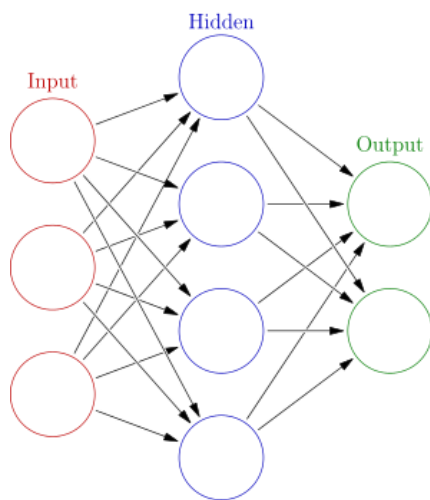


Figure 9 - Basic Artificial Neural Network

The image above describes a simplified fully-connected ANN. Each colored circle represents a **neuron**, and each line represents a connection (i.e. a method of sending information) from one neuron to another, with the arrow showing the direction in which the information travels. Each connection also contains a numerical value called a

weight within it; these weights are used to determine the relative importance of a given connection between neurons. Each weight value is set at random when the ANN is first started up on a computer, generally between 0 and 1. Naturally, randomly-valued weights aren't very helpful at first.

In addition, each column of neurons represents a **layer** of the ANN. There are three layers: the input layer with three neurons, hidden layer with four neurons, and output layer with two neurons. The input and output layer are straightforward – the input layer is made up of the raw input data, with each feature (i.e. distinct column of input data) often given its own input neuron or set of neurons. However, since each input neuron generally will only accept a value between 0 and 1, more complex features either need to be normalized or encoded into multiple neurons. Likewise, the output layer neurons each represent an output for the ANN after each input – sometimes there is only one, sometimes many more. This output is generally in the form of values between 0 and 1. Even in cases where the output appears to be text prediction or image generation, it is often just an amalgamation of outputted values from 0 to 1.

However, the most interesting layer of an ANN is the hidden layer. In the figure above, we can see that each hidden layer neuron receives a connection from all the neurons in the preceding layer (i.e. to its left) and connects to each neuron in the output layer (i.e. to its right). The model above describes a *fully-connected ANN*, so each neuron

is connected to every neuron in the preceding layer and every neuron in the succeeding layer.

The Training Phase

The process below describes a single iteration of the training phase. During the training phase (presuming a supervised machine learning algorithm), a given row of data (i.e. one instance of each input variable's value) moves through the ANN from left to right. First, the input variable(s) become the starting values of the input layer neurons. Then, they are transmitted along the arrow lines to each neuron in next layer to the right (the hidden layer), along with the weight of the connection. From the figure above, since there are three neurons in the input layer (and thus three arrows connected to each hidden layer neuron), each neuron in the hidden layer receives three "value-weight" pairs from the neurons that are connected to it.

Before the hidden layer can send its values to the output layer, it needs to do two steps. First, it needs to generate the new neuron's value by combining the value-weight pairs together. To do this, it multiplies each *value* from the preceding neuron to its associated *weight*, then sums those combined values together. In other words, the new neuron's value becomes:

$$NeuronValue = (v1 * w1) + (v2 * w2) + (v3 * w3).$$

Next, this new neuron's value is inputted into the **activation function**. In other words, with the exception of values going from the input layer to the hidden layer, before any neuron can “fire” (i.e. transmit its information to the neurons connected to it via an arrow) the activation function is applied to the neuron's value.

One common activation function is known as the *sigmoid function*; in the equation below showing the sigmoid function, A is the neuron's output, e is a mathematical constant known as Euler's number (~ 2.72), and z is the neuron's value:

$$A = 1/(1 + e^{-z})$$

Once the activation function is applied, this new value (A) becomes the output of the hidden layer neuron on its way to the neuron(s) in the next layer (in this case the output layer). If there were multiple hidden layers, this process would be repeated for every connection.

However, since the weights were chosen by the computer at random in the beginning, the output for this first piece of training data is going to likely be incorrect, and even if correct it would have been due to luck. Once that determination is made, the ANN then goes “backwards” in its processing. In other words, instead of left-to-right, it now goes from right-to-left, starting with the output layer values. From the connections to those values (i.e. the arrows originating from the hidden layer neurons), the algorithm determines which connection weights were most valuable and which

were least valuable in terms of arriving at the correct answer (or in this case, the incorrect answer). Then, the weights are modified depending on which neurons were deemed most or least responsible for the correct or incorrect output. Slowly, over thousands or even millions of iterations, the weights become more and more accurate. The most common method for this feedback is known as *backpropagation*, although others also exist (Rumelhart, Hinton and Williams, Learning Representations by Back-Propagating Errors 1986).

The Testing Phase

The testing phase is generally similar to the training phase, except with two key differences. First, while the data is still run through row by row, the weights of the ANN don't change. In other words, every step except backpropagation occurs. Second, the purpose of running through the data row by row in the testing phase isn't to improve the accuracy, it's to assess how accurate the ANN has become and whether or not it requires further training to attempt to become still more accurate.

2.2.4 Unique Strengths and Weaknesses of ANNs

ANNs have several critically important strengths and weaknesses to keep in mind. Of note, the strengths and weaknesses noted below are just those which generally differ from other machine learning techniques, or at least appear more strongly for ANNs. Several of them were discussed briefly above and will be covered in \greater depth here.

Key Strengths

The key strengths of an ANN generally include: (1) feature extraction, (2) handling multiple and varied input, (3) high accuracy at a wide range of tasks, and (4) generalizability.

Feature Extraction

First, feature extraction simply means that the ANN can figure out for itself what the important features (i.e. variables) are. In other words, it determines what the patterns for mapping input to output looks like. For example, consider an ANN that judges what a hand-written number should represent on the computer (0-9). No human being ever creates an algorithm which explicitly defines “this is what a 3 looks like”. Rather, the ANN builds its own understanding of what a 3 should look like based on the 3’s labelled in the dataset.

Multiple & Varied Input

Second, ANNs can handle almost all kinds of inputs simultaneously: text, numbers, images, sound, video, and more. The researcher also doesn’t need to worry (as much) about extraneous or irrelevant variables as with other methods – given enough data to learn from and well-tuned hyperparameters, the ANN will naturally reduce the weight connections for unimportant inputs.

Wide Range of Tasks

Third, the sheer flexibility of ANNs is powerful. The layers described above in *Figure 9* are simple, fully-connected layers, but there are many other types of layers which specialize at certain types of tasks, and those layers can be mixed and matched to achieve superior performance for unique problems. Examples of such layers include convolutional neural networks (CNNs) and long short-term memory (LSTM) neural networks, which will be discussed in Chapter Three below. The ability to shape the *structure* of the network can be quite powerful for improving performance.

Generalizability

Finally, ANNs have a great deal of generalizability. For example, with a technique called transfer learning, an ANN trained to recognize one type of image can potentially be used to help recognize another totally different type of image without fundamentally changing the ANN, but merely giving it new input to learn from (Browlee 2017).

Key Weaknesses

However, the weaknesses of ANNs are equally as important to consider, as they have weak points in places where other ML algorithms don't. The core weaknesses of an ANN include: (1) massive training data requirements, (2) advanced hardware requirements, (3) lack of global convergence, (4) substantial technical skill to create, and (5) being a "black box."

Training Data Requirements

The training data requirements for an ANN are often extensive. While all machine learning is considered data hungry to a certain degree, only ANNs have research showing that (in some cases) “...performance increased logarithmically with increasing training data size” (Mitsa 2019). While even ANNs have a certain point at which more data will not improve performance, that point is generally considered to be a significantly higher number than with other types of machine learning.

Advanced Hardware Requirements

While it does not take particularly advanced computer hardware to *run* an ANN on a computer that someone else has trained (i.e. a pre-trained neural network), the training itself requires substantial processing power. Presently, Graphical Processing Units (GPUs) are often used to accelerate this training faster. While smaller ANNs can be trained on a budget ~\$1200 desktop computer with a single good GPU, the largest models today have billions of parameters and can require over 500 GPUs running concurrently to train (T. Ray 2019).

Lack of Global Convergence

ANNs also don't have something known as *global convergence*. What this means is that the ANN you've trained may not be the best solution that the ANN algorithm could achieve. For example, with logistic regression, given the same parameters and the same split between training and testing data, it will output precisely the same model each time it is trained. The reason this does not occur with ANNs is because of their randomized starting weights. If those random weights happen to be too inaccurate, it's

possible that they will never find the best combination of weights to achieve the highest accuracy.⁵ In practice, it is possible to train the same exact ANN structure with the same exact input data on two separate occasions and end up with two entirely different models (with different levels of accuracy) simply because the initial starting weights were different.

Substantial Technical Skill

ANNs also require substantial technical skill to train correctly. There are many parameters that need to be set when determining the structure of the ANN (number of layers, number of neurons, etc.), most of which don't have a clear or universal scientific process for determining what they should be set to. Instead, it can rely on trial and error and intuition. Indeed, many computer science scholars have gone so far as to say that creating an ANN can be "as much art as science" (Chang, et al. 2016).

Being a Black Box

Finally, as has been discussed previously, ANNs are a methodological black box. That is, you can see the data you're inputting, and you can see the result that outputs, but you cannot easily see *why* a given input created a given output. For the simplest of ANNs, looking at the hidden layer's weights can provide a general idea as to what the

⁵ For more on global convergence, see <https://cs.stackexchange.com/questions/2406/must-neural-networks-always-converge>

important variables are. However, this technique becomes vastly less effective for larger and more complex ANNs.

This 'black box' weakness can be crippling when it comes to some kinds of policy analysis, particularly when it is not simply the final prediction accuracy that matters but explaining *why* that result came about. For example, many other kinds of machine learning provide at least a few explanatory statistics. This includes (to name a few): p-values, correlations, statistical significance, confidence intervals, and standard error. In contrast, ANNs provide little usable information besides the raw outputted prediction. While some techniques have been studied to try and "gray" this black box nature of ANNs (many of which will be discussed in the literature review), at present they cannot yet fully compensate for this lack of explanatory information.

2.2.5 Conclusion

This section should by no means be seen as exhaustive of all relevant terminology. Indeed, the 'rabbit hole' of ANN terminology can get deep quickly, with each term sometimes requiring explanations of several other terms before they are understood. Nevertheless, the information provided in this chapter should make the next two chapters easier to understand for the public policy reader.

3 Background

Given that the original artificial neural networks date back all the way to the early 1940s, it may not seem accurate to call them the “newest” of AI techniques. Regardless, because of the scope of time being covered, I use a chronological structure to provide background information for how we arrived at this point in ANN development. In this chapter, I focus primarily on the key events that occurred which shaped ANN/ML development since the 1940s, but with a particular emphasis on the intersection between public agencies and ANN/ML systems.

3.1 Early History: Foundational Papers and Conferences

The creation of the most foundational concepts in “machine” learning itself (i.e. basic linear regression) actually far predates the invention of the computer, or even the typewriter for that matter (Legendre 1805). At the time, regression calculations were done by hand, and so one might argue that linear regression isn’t truly a machine learning technique and shouldn’t be classified as a kind of artificial intelligence, however simplistic.

But aside from these early mathematical innovations, the “father of AI” is generally recognized as British computer scientist Alan Turing. Turing was made famous

for his work to crack the Germans' "Enigma" code machine during World War II that they used to encrypt their communications. Turing and his team's "Bombe" machine was developed to decipher Enigma's encoded messages. In doing so, they became the creators of the first practical AI system (S. Ray 2018). However, it is important to note that while it certainly helped lay the foundations for the field of machine learning, the Bombe machine itself was not actually a machine learning system but rather based on symbolic AI principles.

Artificial neural networks themselves began as a field of study soon afterwards in 1943 with a paper theorizing how biological neurons worked in mathematical terms (McCulloch and Pitts 1943). This initial research was reinforced with the subsequent publication of *Organization of Behavior* (Hebb 1949), which posited that neural pathways are strengthened each time they are used. In the years that followed, computational power expanded to the point where such theories could be tested.

ANN research formally moved into the experimental stage in 1956, with the Dartmouth Summer Research Project on Artificial Intelligence. It brought together researchers from all over the world to discuss the nascent field of artificial intelligence (including neural networks) and pool available research. The conference's mandate was "to proceed on the basis of the conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" (Garson 1998, 3). It was at this conference that Nathaniel

Rochester of IBM attempted to show off the first-ever applied neural network. Although its seemingly nonsensical output at the time wasn't fully understood, this nevertheless amounted to another step forward (McCarthy, et al. 1955).

In fact, it was immediately after this conference that the federal government took its first interest in ANNs. In 1957, Dr. Frank Rosenblatt created the first functional artificial neural network at Cornell Aeronautical Laboratory called the *Perceptron*. While Dr. Rosenblatt himself wasn't an employee of the federal government, his research was nevertheless funded by the US Navy's Office of Naval Research (ONR) (Olazaran 1996, 621).

Because of his success at Cornell, in 1958 the Navy's Weather Bureau employed Dr. Rosenblatt to build a Perceptron neural network for them simply named "704", at a cost of nearly \$17 million 2017 dollars (New York Times 1958). Unfortunately, while this Perceptron technically functioned in the basic experiments they showed off to news reporters (for instance, having it teach itself the difference between *right* and *left*), Dr. Rosenblatt had made vastly overoptimistic claims to the press about the capability of his neural network in the near future. These included claims such as the ability to "walk, talk, see, [and] write" (New York Times 1958). This exaggeration led to substantial controversy from within the emerging field of AI, especially when of course no such ANN was created (Olazaran 1996).

Nevertheless, ONR soon made another foray into ANNs, this time creating the world's first applied neural network with real-world implications. In 1959, ONR funded work for a neural network to eliminate echoes on phone calls. It worked, and ADALINE became the first neural network applied to a real-world problem in 1960 (Widrow and Hoff 1960). However, there is little evidence of continued ONR interest after this point. It was also around this point that the term "machine learning" itself was spread and popularized (Samuel 1959).

Despite ONR's apparent lack of continued interest, however, it appeared ANN researchers had suddenly won the research lottery: a major new federal player had entered the field of computer science research with the creation of the Advanced Research Projects Agency (ARPA) in 1958. ARPA was later renamed DARPA (adding Defense) in 1972 and will be referred to as such throughout the remainder of this paper to avoid confusion. Suddenly, there was a governmental organization set up as a matter of public policy for funding cutting-edge computer science research.

Unfortunately, hopes around DARPA being supportive of ANN research were quickly disabused. Although there was some brief interest in DARPA's first few years, that quickly faded (Anderson and Rosenfeld 1993, 303). This isn't surprising, given (a) the then-ongoing controversy around the Perceptron, and (b) DARPA's close ties to symbolic AI researchers since its founding. Even with the founding of DARPA's Information Processing Techniques Office (IPTO) in 1962, which was set up for the

explicit purpose of creating a new generation of computers which would be able learn and improve over time, DARPA refused to touch ANN research. Rather, almost all their AI funding was focused on symbolic AI research alone (Olazaran 1996, 635-636).

By the middle of the 1960s, it was clear that there were multiple fundamental problems with developing ANNs further: (1) a lack of sufficient computational power, (2) a lack of institutional support/funding anywhere, public sector or private, (3) exaggerated early claims about neural networks leading to resentment and disappointment, and (4) fierce proponents of symbolic AI who genuinely thought ANNs were a lost cause (and if we're being cynical, a potential loss of funding for themselves). These factors combined to create an increasingly hostile research environment for scholars interested in ANNs.

3.2 The First AI Winter

The culmination of these trends resulted in one of the most important pieces of scholarly literature in artificial neural network history: the 1969 book entitled *Perceptrons* (Minsky and Papert 1969) argued that the current design of neural networks was fundamentally unworkable for more complex problems. In more technical terms, Minsky and Papert argued that because the XOR function was not linearly separable, it could not be done with any existing neural network architecture.

This assertion was later to be proven wrong, but at the time it had a substantial impact on the artificial intelligence research community. In fact, although it wasn't officially published until 1969, many drafts were well-travelled within the AI research community during the mid-1960s (Olazaran 1996, 629). Intentionally or not, their book helped bring about the *First AI Winter* for ANN researchers – the federal government wouldn't touch neural network research, and the technology had little real-world application for private industry to get interested in. Many scholars ended up moving to other kinds of machine learning or focusing on the *symbolic AI* branch of artificial intelligence instead.

During the next two decades, there was almost no significant federal or private research support for neural network research. Although a few scattered scholars made occasional contributions and additions (Garson 1998, 5), there was little in the way of sustained advancement in the field.

3.3 Backpropagation and the Thawing of Winter

It wasn't until 1986, with the popularization of the technique of *backpropagation*, that ANN research was finally able to throw off the first AI Winter (Rumelhart, Hinton and Williams, Learning Representations by Back-Propagating Errors 1986). Although these authors were not the first to discover backpropagation (Werbos 1974), which they themselves admitted, their straightforward explanation and

prominent publication venue in the academic journal *Nature* finally spread the technique to scholars across the field. Indeed, the trio followed up their initial publication with a much more in-depth analysis of backpropagation which also addressed the problems addressed by Minsky in *Perceptrons* (Rumelhart, Hinton and Williams, Learning internal representations by error propagation 1986).

This “discovery” of backpropagation helped lead to the end of the First AI Winter, and a variety of publications followed. In 1989, another publication was released that is today almost universally cited as a key stepping stone (Hornik, Stinchcombe and White 1989). The authors’ key contribution (partially funded by the National Science Foundation) was that they “mathematically proved that multiple layers allow neural nets [ANNs] to theoretically implement any function, and certainly XOR” (Kurenkov 2015). Today, these kinds of ANNs are known as multi-layer perceptions (MLPs).

Suddenly, the federal government was interested in neural networks for the first time since the Perceptron in the 1950s (Anderson and Rosenfeld 1993, 299, 306). DARPA reversed itself completely, culminating in their own massive formal study on neural networks. In it, they concluded that “[i]t is time for DARPA to re-examine neural network capabilities.” (Widrow, DARPA Neural Network Study 1989, 52). Indeed, in the years that followed it appears that DARPA made some targeted research investments in ANNs. Also around this time period, the Canadian government became interested in

ANNs - the Canadian Institute For Advanced Research (CIFAR) brought in two noted neural network scholars in 1987, Yann LeCun and Geoff Hinton, which would later prove to be a prescient decision on their part (Bergen and Wagner 2015).

With the widespread dissemination of both backpropagation and the mathematical refutation of Minsky and Papert's assertion through using multiple layers, variations of neural networks began to appear for more specific tasks beyond MLPs (which generally handled numerical data). The invention of convolutional neural networks (CNNs) was a specialized variant that handled visual data. In (LeCun, et al. 1989), the authors created a neural network that analyzed handwritten zip code digits on mail. Prior to their work, computers had had extreme difficulties managing to interpret the subtle differences and imperfections in human writing, even for something as specific as 0 through 9. In fact, the work of (LeCun, et al. 1989) went on to become the "basis of [a] nationally deployed check-reading systems," which was one of the first large-scale implementations of a neural network (Kurenkov 2015).

Yet another ANN innovation following the popularization of backpropagation was how to use neural networks for *unsupervised* learning tasks. Clustering algorithms within the field of machine learning more generally were nothing new, but (Bourlard and Kamp 1988) brought about the idea of *autoencoders* into the ANN mainstream. An autoencoder is an unsupervised (i.e. using unlabeled data) ANN that learns how to compress and encode a particular kind of data and then reverse the process to regain

the original data back (Badr 2019). Soon afterward, ANNs also entered *reinforcement learning*, showing a level of flexibility generally unmatched by other machine learning techniques (Narendra and Parthasarathy 1990). By the start of the 1990s, neural networks seemed to be on fire in terms of research dollars.

For example, DARPA funded a speech recognition research project through SRI International from 1990 to 1997 based on neural networks that produced over a dozen scholarly publications on the subject (Abrash, Cohen and Franco 1997). In addition, they hosted the *DARPA Artificial Neural Network Technology Program Review Conference* in Arlington, VA from at least 1991-1994⁶ (SRI 1997).

While the generally numerically focused multi-layer perceptions (MLPs) and the generally visually-focused convolutional neural networks (CNNs) had been built up until this point, two more ANN variants soon joined the kinds of supervised learning ANNs, these with a focus on interpreting *textual* data. Recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) neural networks were both major innovations during this period on textual data. RNNs were first conceptualized as far back as 1982 (Hopfield 1982), although they didn't gain more major attention until a decade later (Bengio 1993). The LSTM followed in 1997 (Hochreiter and Schmidhuber 1997).

⁶ I could find no official public record of those conferences beyond secondhand accounts, and they may have gone on for additional years beyond this period.

RNNs specialize in processing *sequences* of information, or information where the order of the information (such as sequences of letters making up words) is particularly relevant. Whereas MLPs and CNNs deal with a fixed length for input and output, RNNs can have widely varying lengths. This makes sense considering how one sentence can vary from another in terms of length; this is in contrast to an image as the input, which are generally standardized in terms of the image's dimensions. This makes RNNs widely used in natural language processing tasks (i.e. dealing with textual content) (Zhou 2019). LSTMs are a further variant of an RNN: they modify the algorithm to better “remember” past data. In other words, LSTMs are often used when the most relevant sequence (i.e. word/phrase/sentence) isn't just the current one, but sequences that may have been seen previously as well (Mittal 2019).

However, despite the innovations over this period, the renaissance for ANNs was not to last. ANNs were still incredibly difficult to train and use. The parameters were often difficult to set correctly for training, they could not explain *why* their results were accurate (an ongoing problem to this day), and the amount of processing power required to train them was extensive, particularly in that era of computing. Even with new mathematical learning techniques and increased funding, there just wasn't enough computational power or data available to make truly complex (i.e. deep) ANNs function in the real world for many tasks.

It appears that DARPA also gave up: from October 1998 to October 2009, I could find no publicly available evidence of *any* DARPA funding for ANN-oriented projects.⁷ (Jackson 1998) (Johnson 2009) Simply put, ANN performance was unable to exceed more traditional expert systems computer science techniques at almost any task, and many machine learning scholars had given up on the subfield once again (Allen, How a Toronto professor's research revolutionized artificial intelligence 2015). These culminated in the beginning of the Second AI Winter.

3.4 The Second AI Winter

With the advent of new and powerful machine learning techniques such as random forests (Ko 1995) and support vector machines (Cortes and Vapnik 1995), ANNs soon fell into a rut. While backpropagation had shown itself effective at simple tasks like hand-written digit recognition, it had failed to scale up to larger tasks efficiently with available processing power. These two new techniques, in contrast, proved to be quite effective, faster to train, and easier to use. Indeed, even famed neural network scholar Yann LeCun noted that support vector machines surpassed all but the very best neural networks, while at the same time being substantially easier to use (LeCun, et al. 1995).

⁷ Considering the potentially classified nature of some of DARPA's projects, this may not indicate that no such research occurred, however. Additionally, it is possible that there were obscure projects that escaped my notice during this period.

By the early 2000s, there were estimated to be less than a half dozen artificial neural network specialists worldwide (Allen 2015). However, also around this time, the researchers at CIFAR began to make some interesting discoveries. With CIFAR's support, one of the field's top research scholars George Hinton published two seminal works which caused the beginning of the end for the Second AI Winter (Hinton, Osindero and Teh, A fast learning algorithm for deep belief nets 2006) (Hinton, Osindero and Teh, A fast learning algorithm for deep belief nets 2006). In the latter work, Hinton and his co-authors proved that a neural network could achieve a record-breaking accuracy of 98.75% against the MNIST dataset (a well-known handwritten digit recognition benchmark for machine learning), surpassing the then-record of 98.6% utilizing more traditional ML techniques. To solve the problems of backpropagation, they utilized new advances in processing power, enabling them to build additional layers into their network that weren't previously feasible.

3.5 Taking Machine Learning by Storm: 2006-2013

Hinton and LeCun's work started the revolution that we see ongoing today. Although their work back in 2006 did not initially have too spectacular a response (an improvement of 0.15% isn't particularly grand, after all), what followed unarguably was. Hinton et al's publication began a wave of renewed academic interest. That interest soon flourished into additional advances: by 2009, Hinton and two of his students

developed a neural network that set a then-record for accurate speech recognition, vastly outstripping previous results (Mohamed, Dahl and Hinton 2009).

At long last, the floodgates had broken and ANN research exploded with interest. With more and more scholars entering the field and funding starting to open up from DARPA and private industry, vastly improved predictive accuracy at a wide array of tasks soon poured in: achieving a success rate of 99.65% on the MNIST dataset (Ciresan, et al. 2010) and classifying 1.7 million images into 1000 different categories with a record-breaking accuracy of 84.7% and beating the previous record by over 10% (Krizhevsky, Sutskever and Hinton 2012), just to name a few. Since 2013, well-designed ANNs have generally matched or surpassed other machine learning techniques (in terms of raw accuracy) in most complex tasks.

3.6 Conclusion

Beginning in about 2012 and continuing to accelerate since then, ANN scholarship began to look deeper outside of raw predictive accuracy. While some of the research threads defined in the next chapter had already been well-established studying problems with other kinds of machine learning, it was at this point that they began to flourish for ANNs specifically. Excluding foundational works from general machine learning, this is the time period (2012-2020) where most of the literature review takes place.

4 Literature Review

From their founding as a theoretical concept 1943 up until about 2012, almost all ANN scholarship (and to a somewhat lesser extent machine learning scholarship more broadly) focused on a single, overwhelming objective: *maximizing predictive accuracy*. It isn't hard to understand why – the most fundamental purpose of artificial intelligence is to make decisions, and thus it stands to reason that a system which is able to produce *correct* decisions more often should be superior. What's more, the inherently standardized, quantitative, and comparable nature of accuracy as a measurement of success has allowed scholars to directly compare and compete with one another for whose machine learning system was “best.” Today, we have open-source ML competitions, such as those hosted on Kaggle, where scholars and amateur researchers alike can compete to build the most accurate system for a given task (Kaggle.com 2018).

Starting in about 2012, however, other issues began to percolate to the surface. ANNs were starting to overtake existing ML methods (not to mention traditional approaches based on pure statistics or symbolic AI) in many areas, and it soon became clear that because of their particularly unexplainable and complex internal behavior (among other issues), accuracy alone wasn't enough.

For this literature review, the six research threads I choose to focus on are as follows:

- Accuracy
- Explainability
- Fairness
- Robustness
- Privacy
- Democratic Legitimacy

Within each of these six research threads (perhaps aside from accuracy), there is easily enough scholarship to allow for an entire dissertation's literature review and more. Because of this, I was of necessity highly selective in the literature I covered.

Section 4.7 below also delves into additional threads that were considered but not used as part of the literature review. For each thread, I predominantly include only four types of literature:

- a) Foundational literature which began the research thread under consideration,
- b) Where applicable, literature discussing the key disputes within the thread itself (i.e. when is an explanation sufficient? How is fairness defined? What makes up democratic legitimacy?),
- c) Literature creating a taxonomy of techniques gleaned from a review of previous literature, and
- d) Literature on how to best assess/measure a given research thread (i.e. what indicates when "explainability" rises or falls?)

4.1 Accuracy

Accuracy is the original and (in theory) the simplest research thread to delve into. Merriam-Webster dictionary defines it plainly enough as "conformity to truth or to a standard or model" (Merriam-Webster Dictionary n.d.). However, what most people naturally think of as "accuracy" is really just one specific and intuitive method of calculating this "conformity to truth" for a given mathematical model. Alongside

accuracy, there are also several other potentially relevant metrics that are often found in machine learning scholarship. Most prominently, this includes: recall, precision, F1 score, markedness, and informedness (Powers 2011) (Shung 2018).

To better understand these concepts and their importance relative to accuracy, first let's define a hypothetical ML model that aims to predict whether someone has been diagnosed with a deadly disease. Based on the previous data of one million individuals, we know that the model has an accuracy of 99%. That is, 99 in 100 times it accurately predicts if someone has the deadly disease. If this model were to predict that *you* had this deadly disease, then, should you be worried? The intuitive answer for many people of course is immediately "Yes!", but this isn't *necessarily* the case.

To better determine if you should be worried about the diagnosis, let's add on one additional data point - what if we *also* knew that only one thousand of the one million people in the dataset actually *had* the disease to begin with. Now we can create a *confusion matrix* (Data School 2014) which will help to calculate the five metrics discussed above:

Table 1 - Disease Predictions vs. Reality

	Predicted to have Deadly Disease		
		YES	NO
Actually have Deadly Disease	YES	990 (True Positive, or TP)	10 (False Negative, or FN)
	NO	9,990 (False Positive, or FP)	989,010 (True Negative, or TN)

As we can see above, 99% of people were accurately diagnosed – that includes the $\frac{990}{1000}$ people who actually have the disease and the $\frac{989,010}{999,000}$ people who don't.

4.1.1 True Positive, False Positive, False Negative, and True Negative

The four concepts noted above (also referred to as TP, FP, FN, and TN) make up the confusion matrix, and they are the key values which help us understand what accuracy and these related terms indicate. However, before going any further, lets answer the original question: should someone predicted as likely to have the disease by this model be worried?

With the confusion matrix before us, the answer becomes much clearer. From it, we can see that a total of 10,980 people (TP + FP) were *predicted* to have the disease by the model, regardless of if they *actually* have it. However, we already know that only 990 of them (TP) have the disease – by definition the false positives are just that, false.

By calculating the percentage of actually true cases over all true cases (i.e. $\frac{TP}{TP+FP} =$

$\frac{990}{10980} = \sim 0.0902$), we can see that only about 9% of the people that the model

predicted “yes” for will actually have the disease! This is certainly much less fear-

inducing than the original accuracy number assessed initially. As will be discussed in the

next section, this measurement is quite useful, and is known as *precision*.

4.1.2 Recall, Precision, and F1 Score

The confusion matrix is a critical component for understanding *recall* and *precision*, and from them the F1 score. First popularized in their current form in 1955 for the then-burgeoning field of information retrieval (Kent, et al. 1955), recall and precision are now used widely when analyzing the results of binary classifications (i.e. True/False decisions) from an ML model. Recall (also known as sensitivity) is the likelihood of finding something that is truly there (only looking at true predictions), while precision (also known as positive predictive value) is the likelihood that what you *predict* to be true is *actually* true. These two concepts, along with accuracy itself, are defined mathematically below:

$$Precision = \frac{True\ Positive}{(True\ Positive + False\ Positive)}$$

$$Recall = \frac{True\ Positive}{(True\ Positive + False\ Negative)}$$

$$Accuracy = \frac{True\ Positive + True\ Negative}{(True\ Positive + True\ Negative + False\ Negative + False\ Positive)}$$

While accuracy needs little explanation, precision and recall are meant to capture sensitivity to high false positives and high false negatives respectively. For our case above, let's calculate the precision again (which we've already calculated previously) and the recall:

$$Precision = \frac{TP}{(TP + FP)} = \frac{990}{10,980} = \sim 0.09016 = \sim 9\%$$

$$Recall = \frac{TP}{(TP + FN)} = \frac{990}{1,000} = 0.99 = 99\%$$

In this case, it just so happens that the recall measurement is identical to the accuracy, but this is in no way assured. For the example above, the recall statistic tells us that among those who already have the disease (regardless of whether the model predicted they would), there's a 99% chance that the model would correctly predict it. Thus if you somehow came in knowing you had the disease, the model would almost definitely also say you had the disease. In contrast, the precision tells us that among all the people the model predicted would have the disease, only ~9% actually have it. In this case, clearly precision is the more useful and relevant metric, but that won't always be the case.

Finally, the F1 Score is the "harmonic mean" of the recall and precision measurements (Hayes 2019), and is generally thought of as a better measure of incorrect classifications than accuracy because it penalizes more extreme values. It is defined below:

$$F1\ Score = 2 * \frac{(Recall * Precision)}{(Recall + Precision)} = 2 * \frac{.08926}{1.08016} = \sim 0.16527 = \sim 16.527\%$$

A visualization can also be helpful:

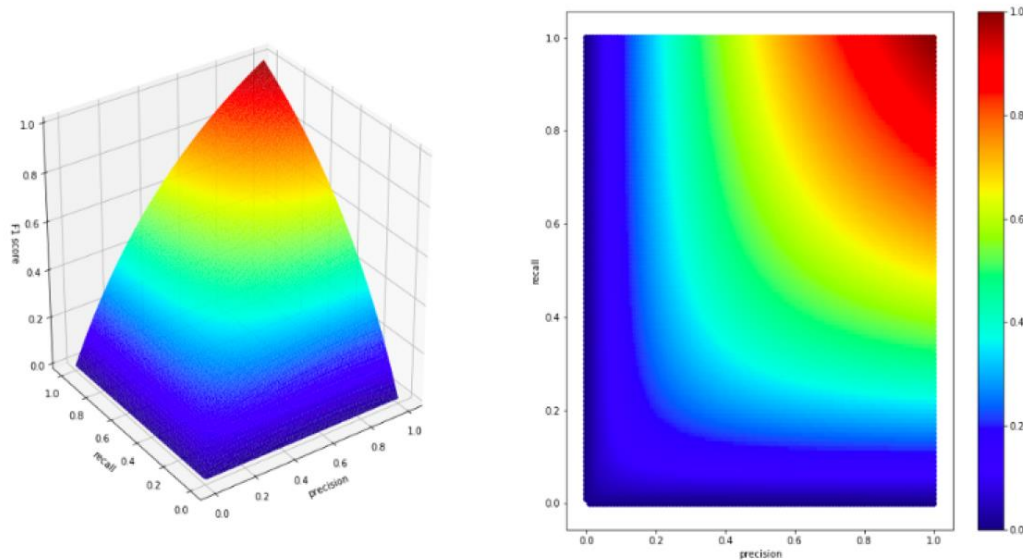


Figure 10 - Changes in F1 Score by Recall/Precisions

Credit: (Mikulski 2019))

We can see from *Figure 10* above that rather than simply averaging the recall and precision together, the F1 Score is more pressured by extremely low values. For example, regardless of how high recall gets, so long as the precision is fairly low the F1 score is going to be fairly low, and vice versa. By contrast, the only way that a high F1 score is going to be calculated is if both recall and precision are both high.

In summation, while a model has 99% accuracy gives it the appearance of being an excellent model, a low F1 Score can show this to be a false veneer.

Criticisms of Recall, Precision, and F1 Score

However, while these other metrics may have advantages over accuracy in some respects, these alternative evaluation methods are not themselves without critics. First, all three values are critiqued because they are entirely insensitive to changes in the True Negative value (Powers 2011, 38). That is, the True Negative value can be changed to any value, no matter how extreme, and recall, precision, and F1 Score will not move in the slightest.

Additionally, the F1 Score has been critiqued because it provides the same weight to recall and precision (Hand and Christen 2018). However, this may not always be ideal: depending on the use case, either recall or precision may be the more important value to have maximized. In the example above, clearly precision is significantly more relevant than recall. However, if the precision and recall values were swapped, the less important recall would still have equal weight and the F1 Score would remain unchanged, even though the situation would be quite different. From these critiques, still more alternative evaluation measurements were born.

4.1.3 Informedness and Markedness

Informedness and markedness were identified by (Powers 2011) as alternative mechanisms of evaluation to recall, precision, and F1 Score. While informedness was a new concept which Powers introduced, markedness was previously popularized for use

in psychology under the term ΔP , or Delta P (Allan 1980). First, let's define the terms mathematically, then more intuitively:

$$\begin{aligned} \text{Informedness} &= \frac{TP}{(TP + FN)} - \frac{FP}{(FP + TN)} = \text{Recall} - \frac{FP}{(FP + TN)} = 0.99 - 0.01 \\ &= 0.98 \end{aligned}$$

$$\begin{aligned} \text{Markedness} &= \frac{TP}{(TP + FP)} - \frac{FN}{(TN + FN)} = \text{Precision} - \frac{FN}{(TN + FN)} = \sim 0.09016 - \sim 0.00001 \\ &= \sim 0.09015 \end{aligned}$$

One of the immediately noticeable aspects is that the informedness/markedness for our test case example are not significantly different than their respective counterparts of recall and precision. This is because of the extremely high True Negative value – if this value were more in line with TP/FP/FN then informedness and markedness would be significantly different. Indeed, this is a core difference with both Informedness and Markedness: they both incorporate the True Negative value, which was a key critique of recall and precision. Additionally, they can be anywhere from -1 to 1, rather than 0 to 1 like recall and precision.

Informedness is just as it sounds: it is a measurement for how *informed* a given model is about the positive and negative values. It takes recall as the starting point, but then penalizes it if the model predictions had too high a percentage of false positives among all actually negative values. Markedness, by contrast, is a measurement of whether the model “marked” the data it needed, whether True or False. It takes

precision as the starting point, but then subtracts from it the percentage of wrong false predictions values among all false predictions (true and false).

4.1.4 Accuracy for ANNs: Alternative Measurements

For some special use cases of ML systems, accuracy itself (including derivations such as F1 Score, etc.) are not used at all. Rather, they have their own highly specialized kinds of pseudo-accuracy equivalents that they primarily utilize instead. These replacements are generally used because the type of task being trained on does not lend itself to easy usage of a traditional accuracy metric.

For example, for an ANN designed to translate from Spanish to English, what is defined as a “correct” answer and what is defined as an “incorrect” answer? This sort of task does not lend itself to easy assessment with the binary options of right vs. wrong given the subjectivity of translations and the ability for an answer to be varying degrees of “somewhat” correct. Because of this, the field of natural language processing (NLP) has invented a wide array of alternative measurements meant to replace accuracy with something more meaningful in their field. The classic example of this is the Bilingual Evaluation Understudy (BLEU) score. First conceived of in 2002 (Papineni, et al. 2002), their paper has since been cited over 10,000 times and is the baseline measurement for a wide range of NLP tasks to this day.

4.2 Explainability

Even with those alternatives to potentially better evaluate a model, accuracy and its derivative metrics alone aren't enough in public policy. Getting to the right answer isn't the only important quality an ANN might have - rather, the *why* can be equally as important, if not more so. Except in the simplest of one-layer cases, artificial neural networks do not natively provide much of any explanatory information. Indeed, the only information natively produced by most ANNs are the final weights of each connection and the raw accuracy of its predictions with those weights. It is from these limitations that the research thread of *explainability*, sometimes also referred to as explanatory power or explainable AI, was born.

It should be self-evident why explainability is highly prized from a public policy perspective: at a minimum, government systems generally necessitate at least some level of transparency and accountability in their processes, but transparency and accountability are meaningless if the public agency *itself* doesn't understand why an ANN made a given decision. Even more fundamental, however, is the question of what exactly explainability even *is* within the context of ANNs, and how “much” explainability is enough. What kinds of techniques exist and how do they differ from one another? Indeed, the level of explainability required even varies by context.

Consider two potential cases of ANN use in public policy: handwritten character recognition and federal loan guarantees. In the former case, explainability needs are

likely to be fairly low: so long as the model is accurate, policymakers are unlikely to be interested in explanations on which curves and corners of the writing make the ANN detect which characters. However, in the latter case we would expect explainability needs to be substantially higher: precisely how the ANN came to a given loan decision should be important in almost any public policy context.

This section primarily focuses on the taxonomy of (Gilpin, et al. 2018) for understanding the different types of explainability techniques that presently exist; Gilpin's work goes into a deeper technical discussion of these issues for those interested. Their paper was chosen for several of reasons. First, the authors provide an efficient taxonomy which covers a broad range of techniques that presently exist and can likely be used to classify many future techniques. Second, many techniques within their taxonomy can be applied broadly to various kinds of ANNs, which was a limitation of other explainability taxonomies considered such as (Grun, et al. 2016). Finally, the authors focus on one of the central purposes of this section: to better define what, precisely, explainability even *means*. It is only with a clear definition that we can compare explanatory power to other research threads.

4.2.1 Foundational Works

Scholarship into explainability for ANNs began in 2013 when (Zeiler and Fergus 2013) introduced a technique to help understand why a convolutional neural network (if we recall, a type of ANN generally used with images as inputs) functions as it does. Their

technique was simple: first, they placed a gray box which covered up some portion of the input image and then had the CNN classify the image. Then, they moved the box along the image and continually had the CNN reclassify the modified image to look at how the output changed. This process was repeated until they had predictions for each possible position the gray box could be in.

In this way, they could attempt to determine which pixels or group of pixels were the most important in the image based on which gray box position caused the greatest change in output. Their work was soon followed by (Simonyan, Vedaldi and Zisserman 2013), who hypothesized that instead of identifying the pixels which caused the most neurons to fire, it is the pixels that require the least change to cause the greatest impact on classification that are most important. From these beginnings, other scholars joined to help understand not only the what, but the why; as of December 2019, Zeiler and Fergus' work has been cited almost 7,900 times according to Google Scholar.

4.2.2 What Makes a Good Explanation?

The concept of an explanation itself has deep roots in philosophy. (Gilpin, et al. 2018) argue that an explanation is sufficient for the purposes of an ANN when there are no further "why questions" that need to be asked. The authors further assert that there are two ways an explanation can be evaluated: by *interpretability* and by *completeness*. Interpretability is how easily a given explanation can be understood by humans, whereas "[a]n explanation is more complete when it allows the behavior of the system

to be anticipated in more situations” (Gilpin, et al. 2018). Thus perfect completeness would have each and every mathematical operation spelled out, whereas perfect interpretability would allow any user (including non-experts) to understand precisely what is being presented and how the presented conclusions were reached. By default, perfect completeness and almost no interpretability is already present; an ANN is just a series of complex number matrices and equations, after all. However, given the millions or billions of mathematical operations which occur to train an ANN, no human could be expected to interpret such a “perfectly complete” explanation, if it could even be called that.

Interpretability and completeness, then, are often in conflict: human beings need some mechanism to simplify a perfectly complete ANN’s explanation into something we can interpret and draw reasonable conclusions from. However, going to the opposite extreme is also potentially flawed: an overly simplified explainability mechanism that merely outputs “yes” or “no” could in some situations be considered incomplete and end up hiding or misrepresenting a significant amount of information. The core issue in explainability, then, is what the proper balance between interpretability and completeness is and how to best reach that balance.

4.2.3 Taxonomy of Explainability Techniques

Rather than attempt to define what that perfect relationship is (a task which is highly subjective, context-dependent, and ever-changing), (Gilpin, et al. 2018) provide a

taxonomy of explainability methods as well as how to evaluate what those methods produce. They split existing techniques into three categories based on how they attempt to explain the ANN in question: processing, representation, and explanation producing techniques.

First, *processing techniques* focus on reducing the complexity of the data within an ANN to the point where it can be interpreted by a human being. Examples include generating graphical visualizations of different weights and activation functions being used, as well as decomposing the ANN into a decision tree and then interpreting the tree's outputs. Second, *representation techniques* focus on understanding how data flows through key structural elements of the ANN, such as its layers, its neurons, or even the general vector direction of its output. Finally, *explanation producing techniques* attempt to create a fundamentally more explainable ANN from the beginning through the structure of the model itself.

4.3 Fairness

Perhaps no topic is more important in the realm of public policy than that of fairness and bias. However, there are many ways to define fairness, some of which violate other definitions. This can make determining what is fair quite difficult, and indeed subjective, depending on the fairness standard one uses.

This section in the literature review will be devoted to *algorithmic* fairness, or in other words fairness as it relates to the choice of algorithms and/or statistics used to

calculate the fairness of an ANN or machine learning system. Algorithmic fairness and bias are essentially antonyms in the context of artificial neural networks and machine learning: an ANN that has achieved perfect algorithmic fairness in a particular dimension (i.e. gender, race, etc.) is not biased along that dimension, and a biased system cannot be perfectly fair. Any further references to fairness in this dissertation should be construed as referring to *algorithmic fairness* unless otherwise specified.

In contrast to algorithmic fairness, there is also *non-algorithmic fairness*, which I define as fairness as it relates to a public agency's structure and decisions more broadly. Non-algorithmic fairness is encompassed within the democratic legitimacy section of the literature review below.

Issues of fairness in ANNs also have a broad applicability to society at large. For example, a 2018 study from Harvard's Center for Internet and Society identified five real-world examples where the potential for bias in AI could be particularly harmful – (1) calculating credit scores, (2) healthcare diagnostics, (3) online content moderation, (4) recruitment and hiring, and (5) automated essay scoring (Raso, et al. 2018). Many of these issues are not directly in the realm of public policy or public agencies, of course – the US government does not calculate an individual's credit score, for example. Nevertheless, the problems achieving fairness faced by private sector actors are often the same.

However, even the most fundamental questions in this research thread are subject to fierce debate. For example, what precisely do fairness and bias *mean* when they must be defined mathematically? This is important, because the only way to implement any kind of “fairness standard” in an algorithm, ANN or otherwise, is to have it defined mathematically.

4.3.1 Foundational Works

Fairness literature into machine learning generally was first introduced with (Kamiran and Calders 2009), and was quickly followed with other works as the field expanded (Pedreschi, Ruggieri and Turini 2009) (Calders and Verwer 2010). However, most ML techniques prior to ANNs were not capable of handling the same kinds of datasets that ANNs now can. For example, an ANN can be structured to simultaneously accept visual and textual data or to perform both supervised and unsupervised learning in the same model; this can require some distinct literature. The foundational work of fairness literature focusing on ANNs specifically is often thought of as going back to 2016, with (Bolukbasi, et al. 2016) analyzing the biases within Google’s “Word2Vec” word embedding ANN.

4.3.2 Conceptions of Fairness: Spaces, Beliefs, and Mechanisms

One of the most thorough and methodical attempts to define algorithmic fairness for the purposes of machine learning is (Friedler, Scheidegger and Venkatasubramanian 2016). The authors begin by summarizing the central defining problem: the inherent tension in many of the existing competing fairness definitions

between *equality of outcomes* (i.e. fairness in result) and *equality of treatment* (i.e. fairness in procedure).

Interestingly, this dispute should already be familiar to many of those who deal with public policy or economics. This is because these two competing definitions of fairness broadly make up the key dividing line between the US political left and right on economics: left-wing economic policy is often more focused on equality of outcome (Matthews 2015), whereas right-wing economic policy is often more focused on equality of treatment (sometimes alternatively referred to as equality of opportunity) (FREOPP 2016).

For an example in public policy today, consider the case of taxation policy – one example of a taxation policy tilted towards equality of treatment would be a flat tax, or in other words a tax system where everyone pays the same percentage of their income in taxes. No matter how much money you make (i.e. the outcome), you still pay the same percentage in taxes. In contrast, a tax policy with multiple tax brackets depending on your income is more closely aligned with equality of outcome – some individuals will pay a higher percentage of their income in taxes than others (which is not equal treatment) in order to attempt to achieve greater fairness in the outcome, such as through using the excess revenue to fund public services.

For their own part, (Friedler, Scheidegger and Venkatasubramanian 2016) provide a taxonomy for understanding different definitions of fairness in the context of ANNs through three overlapping concepts: spaces, beliefs, and mechanisms.

Spaces

The authors first define three types of “space” that overlap with one another to varying degrees: the *decision space*, the *construct space*, and the *observed space*. These spaces are particularly important, as “it is the conflation of these spaces that leads to much of the confusion and disagreement in the literature on algorithmic fairness” (Friedler, Scheidegger and Venkatasubramanian 2016, 2).

The decision space refers to the potential problem that must be dealt with; some examples include employee productivity, the tax rate, prison recidivism, or college admissions. Within that decision space, there is the *construct space*, the central concept within that decision space that is being measured. For example, within the decision space of college admissions, examples of a construct space might include an applicant’s intelligence or their success in high school.

We use the construct space to define the question we want to solve within the decision space. However, the authors are quick to note that construct space is generally impossible to *perfectly* map, no matter how much data we have – after all, there is no such thing as a perfect measurement of intelligence or a perfect measurement of “success” in high school.

Nevertheless, whichever (inherently imperfect) measurement ends up being used within the construct space is the *observed space*. That is, the precise variable(s) used for measurement in the construct space. Whereas the construct space is impossible to perfectly observe, this is not the case for the observable space, which we

can measure and gather data on. Thus, to measure intelligence, one potential observed space variable might be IQ, or to measure success in high school, the observed space might be a student's GPA. However, literature abounds showing that IQ is not a perfect measurement of intelligence (Martschenko 2017) and that student success in high school is not necessarily the same as a student's GPA (York, Gibson and Rankin 2015). Despite this inherent imperfection in measurement, however, the observed space is all we can gather data in.

With the three overlapping spaces in mind, the authors then define two other key concepts: that of *beliefs* (what is believed about the state of our world) and that of *mechanisms* (what methods should be instituted to achieve a belief's ideal).

Beliefs

There are two central beliefs defined by (Friedler, Scheidegger and Venkatasubramanian 2016) that are linked to what we assume about the relationship between the constructed and the observed space: what they colloquially refer to as "what-you-see-is-what-you-get," or *WYSIWYG*, and "we-are-all-equal," or *WAE*. The *WYSIWYG* belief asserts first and foremost that while we may not be able to perfectly map the construct space from the observed space, it is generally similar enough to the observed space such that there is little problem in using measurements for the observed space as a stand-in for the constructed space. Thus, a *WYSIWYG* belief might assert that IQ (the observed space) is a close enough mapping of an individual's intelligence (the

construct space) to use it in the decision space without any further modifications to achieve fairness.

In contrast to this is WAE, which asserts that since we cannot get a truly accurate image of the construct space from an inherently imperfect unit of measurement in the observed space, we should instead begin with the assumption that all relevant sub-groups being measured within the observed space *should* on average have an equal outcome. In those cases where the data doesn't reflect this, the fault then lies either in the observed space being poorly mapped to the constructed space, or the fact that those inequalities in outcome are due to issues of structural bias in one's society or environment that individuals cannot control for and thus should not be held against them mathematically.

In short, the WYSIWYG belief starts with the observed space and then claims the construct space is very similar, whereas the WAE belief starts with an assumption of equality between groups in the construct space, and then claims that differences in the observed space are therefore due to an imperfect mapping between the spaces or structural biases in society.

In comparing these two beliefs (alternatively called axioms), the authors conclude that "[w]hatever the motivation (which is ultimately mathematically irrelevant), the choice in axiom is critical to a decision-making process. The chosen axiom determines what fairness means by giving enough structure to the construct space or the mapping between the construct space and observed space to enforce

fairness despite a lack of knowledge of the construct space” (Friedler, Scheidegger and Venkatasubramanian 2016, 9). Depending on which axiom an individual or public agency believes in, then, will fundamentally shape how they determine what achieving fairness looks like.

Mechanisms

The authors finally define two mechanisms to achieve what they call fairness in result (i.e. equality of outcome) and/or equality of treatment): the *individual fairness mechanism* and the *group fairness mechanism*. An individual fairness mechanism is what we would think of intuitively as fairness at the individual level; it looks at how each individual performs, and the model that makes the correct decision for the greatest percentage of individuals is considered the fairest. Thus if a model gets the answer right for 95% of individuals (i.e. a 5% error rate), that 95% is what an individual fairness mechanism would assess to determine how “fair” the ML system is. Under the individual fairness mechanism, any ML system that obtained lower than 95% correct would be less fair since it got more individuals wrong.

A group fairness mechanism, in contrast, focuses on comparing the results of sub-groups of individuals within the overall group. For example, a group fairness mechanism would consider if poor individuals and rich individuals as respective groups had similar error rates: if rich individuals had an error rate of 2% but poor individuals had an error rate of 10%, a group fairness mechanism would consider that gap as highly

relevant to assessing fairness (although there is no definitive clear-cut numerical metric like there is for the individual fairness mechanism).

These two mechanisms are often at odds with one another – is it better to have a lower overall error rate for all individuals with unequal error rates between different groups (i.e. the rich and the poor), or is it better to have a somewhat higher overall error rate but have equal error rates between the two groups? Is there a certain quantitative error gap where either the individual or group fairness mechanism becomes “better”? Attempting to quantify what such a number should be is a task I don’t envy.

The authors conclude that there is no “magic bullet” for fairness and non-discrimination to both be achieved. Depending on which belief the decision-maker subscribes to, there are different guarantees. As the authors put it, “...under the WYSIWYG worldview fairness [in result] can be guaranteed, while under a structural bias [WAE] worldview non-discrimination [fairness in treatment] can be guaranteed.” (Friedler, Scheidegger and Venkatasubramanian 2016, 12). The authors also assert that the choice of mechanism is equally as important: fairness can only be guaranteed using both the WYSIWYG axiom and the individual fairness mechanism, and non-discrimination can only be guaranteed using both the WAE axiom and the group fairness mechanism. This is not to say that fairness and non-discrimination are not possible otherwise, simply that they cannot be mathematically guaranteed.

In summation:

Table 2 - Fairness Mechanisms vs. Worldview

	Individual Fairness Mechanism	Group Fairness Mechanism
WYSIWYG	Guarantees Individual Fairness in Result	Guarantees Lack of Fairness
WAE	Guarantees Discrimination	Guarantees Group Non-Discrimination

4.3.3 Conceptions of Fairness: Parity, Equality of Odds, and Calibration

However, (Friedler, Scheidegger and Venkatasubramanian 2016) isn't the only scholarship looking to understand and conceptualize algorithmic fairness. An alternative taxonomy and definitions are provided by (Wadsworth, Vera and Piech 2018). Their study focuses more closely on issues of criminal justice and incarceration between blacks and whites. Rather than looking at the WAE vs WYSIWYG dichotomy, Wadsworth et al. assert that "...if black people are more likely to become incarcerated in the US than white people when controlling for criminal behavior...black inmates should not be punished for our biases with harsher recidivism predictions." To put this into the language of (Friedler, Scheidegger and Venkatasubramanian 2016), Wadsworth et al. do not accept that the constructed space (incarceration) is near-identical to the decision space (guilt or innocence judgment); rather, they assert that the inequalities that exist are due to structural issues in society and that algorithms should not punish black

inmates for this. In short, the worldview of (Wadsworth, Vera and Piech 2018) is clearly WAE, and they reject WYSIWYG as an option. Instead, they go into greater depth within WAE and compare different mechanisms of group fairness.

(Wadsworth, Vera and Piech 2018) explicitly focus on the intersection of machine learning algorithms with recidivism cases, specifically the controversy that erupted from the case of COMPAS, a machine learning tool used in the Wisconsin judicial system. COMPAS was designed to determine who is a likely risk of recidivism by inmates. However, in May 2016 ProPublica produced a study asserting that the COMPAS algorithm was racially biased (Angwin, et al. 2016). Six weeks later, research scholars at Northpointe (the firm that makes COMPAS) put out their own competing study claiming that no such bias existed (Dieterich, Mendoza and Brennan 2016).

Wadsworth et al. wade into this conflict by creating their own taxonomy of fairness mechanisms and applying them to the case of recidivism. First, they describe their three definitions of fairness (which should be noted are all different definitions of group fairness, having rejected simple individual fairness as insufficient):

Parity: "...the proportion of individuals classified as high-risk is the same for each demographic" (Wadsworth, Vera and Piech 2018)

Equality of Odds: "...the proportion of individuals classified as high-risk is the same for each demographic, when true future recidivism is held constant. White and black inmates that do recidivate should have the same proportion of high risk classification." (Wadsworth, Vera and Piech 2018) In other words, equality of odds is equivalent to having the same true positive and true negative rates between groups.

Calibration: "...reflects the same likelihood of recidivism irrespective of the individual's demographic. In this application, black inmates who are classified as high risk should

have the same probability of true recidivism as white inmates classified as high risk.” (Wadsworth, Vera and Piech 2018)

The easiest way to compare their definitions is that Wadsworth et al. goes deeper than (Friedler, Scheidegger and Venkatasubramanian 2016) in how far along they dive into group fairness. In fact, parity is generally equivalent to the Group Fairness Mechanism defined by (Friedler, Scheidegger and Venkatasubramanian 2016). However, (Wadsworth, Vera and Piech 2018) go even further with their two additional mechanisms. By describing two additional mechanisms, they are implicitly asserting that parity on its own is an insufficient definition for group fairness.

The next definition, equality of odds, can also be described as true negative plus true positive equality. In other words, among only those inmates that *did* end up recidivating, is parity still true? Finally, calibration flips equality of odds on its head – instead of holding recidivism constant and checking for whether the groups are classified as high-risk equally, calibration holds the original high-risk assessment constant and checks for whether the groups recidivate equally.

However, (Wadsworth, Vera and Piech 2018) found that all three of their group fairness mechanisms could not be maximized simultaneously. When they rebuilt the COMPAS ML algorithm from part of the original data that COMPAS used (the data was released publicly, but the actual ML system COMPAS built was not), the AUC⁸ for their *unconstrained* ANN (that is, when they applied none of their three fairness constraints

⁸ AUC stands for “Area Under Curve”. For more information on AUC as a statistical measurement, please see <https://analyse-it.com/docs/user-guide/diagnosticperformance/auc>

and allowed accuracy to maximize) was 72%. In contrast, their chosen *constrained* ANN (where the gap between different racial sub-groups was 2% or less) had an AUC of only 70%. (Wadsworth, Vera and Piech 2018, 3) The authors also admit that they did worse at calibration than COMPAS' entirely unconstrained approach because they optimized for parity and equality of odds.

Returning to the specifics of COMPAS, although ProPublica's and Northpointe's methodologies were both disputed by other scholars (Flores, Bechtel and Lowenkamp 2016), their most fundamental disagreement (whether or not COMPAS was biased) can be explained with the terminology provided by (Wadsworth, Vera and Piech 2018): Northpointe assessed *calibration* as its definition of fairness, while ProPublica assessed *parity* and *equality of odds* as its definition of fairness. By each of their own definitions, they were correct. What's more, were Northpointe to attempt to achieve parity and equality of odds, they would likely have to damage calibration to do it.

4.3.4 Taxonomy of Fairness Techniques

While the fairness mechanisms discussed above are techniques in the abstract sense, and (Wadsworth, Vera and Piech 2018) even provide a few examples of practical techniques, neither discusses in depth the ways to create these fairness mechanisms algorithmically. From my review of current literature, there are four families of techniques in this taxonomy: data augmentation, preprocessing algorithms, algorithm

modifications, and postprocessing techniques (S. Friedler, et al. 2018) (Chen, Johansson and Sontag 2018).

Data Augmentation

Data augmentation simply involves adding more training data. For example, (Chen, Johansson and Sontag 2018) argue that since most other fairness optimization techniques end up reducing the overall predictive accuracy of their model in search of fairness, they are inherently inferior, especially for critical areas where accuracy is paramount such as healthcare or criminal justice. The authors then suggest that as an alternative, “...[additional] data collection is often a means to reduce discrimination without sacrificing accuracy” (Chen, Johansson and Sontag 2018). However, it’s arguable whether or not this is a generally applicable technique for increasing fairness since there are many situations when more data cannot be obtained.

Preprocessing Algorithms

Preprocessing algorithms are based on the idea that “training data is the cause of the discrimination that a machine learning algorithm might learn, and so modifying it can keep a learning algorithm trained on it from discriminating.” (S. Friedler, et al. 2018). Generally, such techniques don’t involve changing data labels (i.e. the ground truth for the training data), but rather focus on various input modifications (Feldman, et al. 2015).

Algorithm Modification

Algorithm modifications involve model-specific changes to how learning functions in order to reduce or eliminate bias. For example, the ANN produced by

(Wadsworth, Vera and Piech 2018) clearly fits this definition: by applying adversarial learning techniques to their network while it was training, they attempted to achieve their fairness standard. These algorithm modification techniques are often referred to as constrained optimization, or in other words, “obtain the highest possible accuracy during training while not violating a given standard of fairness”. Because the decision boundaries for the ANN are inherently more limited, however, that accuracy is likely to be at least slightly lower than it would be in an unconstrained training environment.

Postprocessing Techniques

Finally, postprocessing techniques function after training has been completed. These techniques modify the output in some way to ensure that a fairness standard is met. An example of postprocessing is (Hardt, et al. 2016), who show how to adjust a learned predictor’s output to remove discrimination. However, their model relies on their specific definition of fairness, as well as their assumption that data about the predictor, target, and membership in the “protected group” are all available in the data.

4.4 Robustness

In this study, robustness is defined as how resistant an ML algorithm is to maliciously manipulated data, including both data poisoning during training and adversarial examples on a trained model. Both issues are potential threats not only to ANNs, but to almost any kind of machine learning system.

Data poisoning is a simple concept – a malicious user provides an ML model in the midst of training with artificially modified data so that the model learns the wrong inferences from the data (Moisejevs 2019). Adversarial examples are similar – they occur not when the model is training, but when the model is done training and an adversary wants to trick the model into misclassifying an input.

Of note, robustness should not be confused with accuracy: although there is often a correlation between the two, a model can be both non-robust and have a high accuracy. This is because unless special procedures are utilized to mitigate the issue, the accuracy of the model is generally based on “normal” instances of the input and is tied to the specific test dataset it is evaluated on (Madry, et al. 2019). However, an adversarial example is maliciously created to fall outside of the normal conditions a model is designed to handle.

For example, consider a hypothetical ML system with 99% accuracy. This 99% accuracy should not be construed as representing “99% accuracy no matter what kind of input is attempted.” Rather, this is 99% accuracy based on the data it was trained on (which, unless it was augmented in some way, is representative of reality only). Because an adversarial example can be subtly manipulated into an input that cannot exist in the real world, accuracy is often substantially lower when faced against adversarial examples versus normal data (Madry, et al. 2019).

The best way to conceptualize an adversarial example is visually, because it provides a striking case. The basic procedures are straightforward to create an adversarial example in the first place. First, begin with an ANN which assesses what kind of object is being displayed in an image. Then, add a specially designed *perturbation* to the pixels in the image and watch as the ANN suddenly asserts the image to be something nonsensical. An actual real-world example is below:

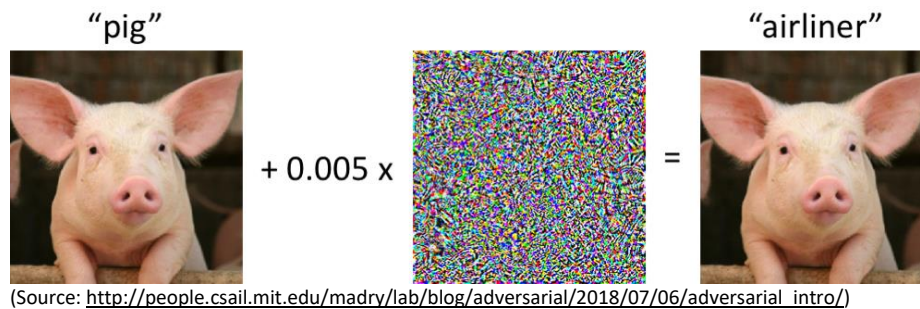


Figure 11 - Adversarial perturbation of image

The initial image on the left is correctly classified as a pig by the ANN. However, by applying a 0.5% change to each pixel based on the static-like image's pixels, the image becomes classified as an airliner, which is obviously incorrect. Of course, no human would make such a mistake – the image that the ANN classified as an “airliner” still looks entirely like a pig. Indeed, the image needs to be observed closely and zoomed in on before a human can even *detect* the slight perturbation in the image with the naked eye. However, ANNs can be fooled by such perturbations. An adversarial example

is, for all intents and purposes, an attack designed to have the system explicitly misclassify the image under consideration.

4.4.1 Foundational Works

Issues of robustness were known in the world of machine learning far before ANNs became prominent. The concept of “attacking” trained machine learning classifiers in such a manner was first published in 2004 against a simpler form of machine learning (Dalvi, et al. 2004). Their article references the ways in which email spammers would defeat rudimentary anti-spam machine learning classifiers (i.e. adversarial examples) and how those classifiers needed to be constantly rebuilt to ensure that their accuracy didn’t degrade too quickly. More scholarship on the subject soon followed (Globerson and Roweis 2006).

Since then, literature on the subject has continued to expand. However, it wasn’t until 2013 that ANNs had achieved enough notoriety in image classification that scholarship started appearing about the problem (Szegedy, et al. 2013). Since then, innumerable papers have been published on the topic. Robustness literature frequently revolves around a “cat-and-mouse” game where new techniques are discovered to make an ANN more robust, only for future literature to then poke holes in those techniques and vice versa.

As just one example, the process of *distillation* of an ANN was originally conceived of to increase its accuracy and performance (Hinton, Vinyals and Anddean

2015). However, some scholars soon took this concept a step further and conceived of *defensive distillation*, or in other words, utilizing the techniques of distillation to make an ANN not only more accurate but also more robust (Papernot, McDaneil, et al. 2016). This new technique quickly became the subject intense debate over its effectiveness, with both attacks against it (Carlini and Wagner, Defensive Distillation is Not Robust to Adversarial Examples 2016) (Carlini and Wagner 2017) and further refinements to it. (Papernot and McDaniel 2017) Other examples of robustness techniques which have been frequent targets of both attack and refinement include obfuscated gradients (Athalye, Carlini and Wagner 2018) and ensemble defenses (He, et al. 2017), among still more.

4.4.2 Robustness Certification Standards

While there is little in the way of taxonomies of robustness techniques, the remainder of this section is focused on analyzing the different methods of *measuring* robustness and what those methods entail.

Arguably the most common method of measuring robustness (as of December 2019) is via something known as a *robustness certification standard*, which states that for a given classification, the certification provides an absolute assurance within a given area of decision space nearby that adversarial perturbations will not cause a misclassification. While there are a great many techniques to try and *find* the largest certified robustness region, most scholarship presently agrees that some kind of

robustness certification is the best standard to measure robustness by. The idea should be made easier to understand with the help of a visual aid:

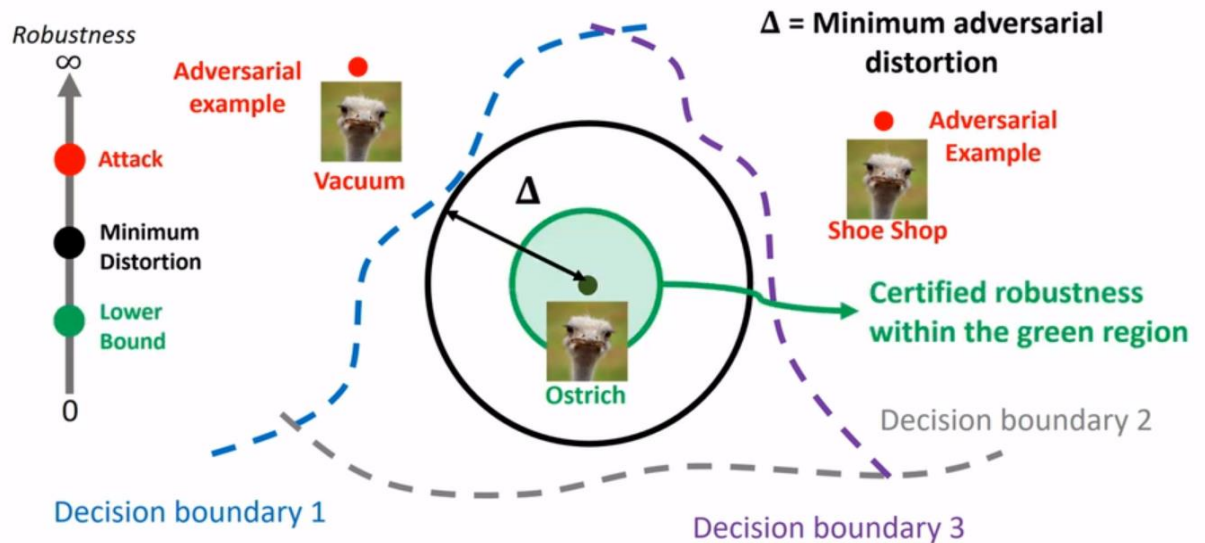


Figure 12 - Robustness Certification

Credit: (Boopathy, Presentation on an efficient computation framework of a certified robustness measure for convolutional neural networks 2019)

The graphic above helps to visualize the idea of robustness certification. First, it simplifies our decision space to only two dimensions, X and Y. Then, we have the classification of this Ostrich image, represented by the center point. The three decision boundary lines show us the actual decision boundary lines within our two dimensional space – if the (X,Y) coordinate for classifying an Ostrich were adversarially modified,

with examples shown by the red dots, the decision boundary shows the limits to which the ANN would still classify the ostrich as an ostrich. Thus, the red dots are successful adversarial attacks in this case.

In this same vein, the black circle shows us the largest circle (since this is simplified to only two dimensions) we can create while still having an ostrich correctly classified as an ostrich. If the black circle were to have its radius increased any more, at least one part of it would fall outside the decision boundary. However, there's an important problem here – while a computer might easily be able to make such a determination in a two-dimensional space, determining the *actual* decision space boundaries (and thus the minimum adversarial distortion) for a 300-dimension decision space (far more common in image classification) is often computationally infeasible. Because of this, a variety of different methods for determining a certified robustness area within the minimum boundary that we *can* calculate have been devised.

While some scholars focused on providing an absolute certification for robustness based on the worst case possibility for an adversarial attack (Zhang, et al. 2018) (Weng, Zhang, et al. 2018) (Boopathy, Weng, et al. 2019), others focused on methods which provide a *nearly* absolute assurance but have significant gains in terms of computational speed. (Mangal, Nori and Orso 2019) (Weng, Chen, et al. 2019). For a deeper technical review of available literature in these two areas, see (Singh, et al. 2019).

In short, while an absolute robustness certification would provide 100% certification within the green circle of *Figure 12* that adversarial examples wouldn't work, the near-absolute certification would provide a lesser degree of surety. As discussed above, however, the trade-off is in computational speed. Different robustness certification techniques, absolute or otherwise, compete on several criteria:

Speed of Computation: Since computational efficiency is one of the key limitations towards calculating this certified area, the mathematical algorithm used to certify the space as robust matters significantly. Even if all algorithms successfully certify a region, an algorithm which can do so 10x or 100x or 1000x times as fast is naturally going to be superior for practical usage.

Flexibility of Application: There are numerous kinds of activation functions and layers that exist with ANNs, and each function differently. An algorithm which can successfully create a certified robustness region for a convolutional neural network, for example, isn't necessarily able to do so on a fully connected neural network.

Absolute vs. Approximate Certification: Absolute certification techniques have the simplicity of simply being able to say "never" when it comes to adversarial examples within their certified area. However, they are also generally computationally infeasible for larger/deeper ANNs as of 2019. In contrast, approximate certifications can certify larger areas for deeper networks, but those robustness certifications have caveats.

Sometimes those caveats are large and sometimes they're small, but they can make assessing robustness more complicated.

Size of Certified Robustness Region: Finally, techniques compete on the size of the overall certified robustness region. A larger certified region means that the network is better able to deal with adversarial attacks and avoid misclassification. Notably, the techniques listed in this section are generally not for directly *optimizing* an ANN to become more robust. Rather, they simply spell out a mechanism to *measure* precisely how robust they are.

4.5 Privacy

Privacy has come to the forefront of ANN research in recent years. Any kind of machine learning, but especially ANNs, take in a massive quantity of data to train with. However, even if the training data itself is anonymized, there is still the threat of de-anonymization (Lee, et al. 2017). That is, even when a given dataset has been explicitly relieved of its individually identifying characteristics, de-anonymization would get that data to be re-linked to the individuals associated with it. Within this research thread, there many methods for creating a “privacy-preserving ANN.” This study will primarily focus on differential privacy (DP), presently the most published technique, although other methods such as homomorphic encryption and federated learning are reviewed as well.

4.5.1 Foundational Works: Differential Privacy

Research into the idea of differential privacy itself began in 2006 with (Dwork 2006), although it was not directly applied to the field of ANNs until 2015 (Shokri and Shmatikov 2015). In their work, the authors identify three key objectives for a privacy-preserving ANN:

- Protecting privacy of the training data
- Enabling participants to control the learning objective and how much to reveal about the model
- Allowing the application of new inputs into a given model without revealing the original inputs or the outputs.

While homomorphic encryption and federated learning are also concepts worthy of discussion, neither has reached the critical mass of ANN scholarship that DP has (Mancuso 2019).

Differential privacy can generally be understood as a set of techniques meant to counteract the de-anonymization of data. There are a wide variety of algorithms which attempt to produce differential privacy, as well as different standards for when differential privacy has been achieved. One easy way to conceptualize DP without math is to consider the case of two near-identical databases, database A and database B. The only difference between database A and database B is that database A contains your information and database B doesn't (i.e. one row of data missing). From there, "[d]ifferential [p]rivacy ensures that the probability that a statistical query will produce

a given result is (nearly) the same whether it's conducted on the first or second database." (Green 2016).

With this basic example, DP would be defeated simply by querying a count of the number of rows – if the database has your information, the result would be one higher. However, such queries (and others like them) can be defended against by adding a small amount of random statistical noise to the result. In other words, by adding a slight amount of *imprecision* to the query result (i.e. if the query returned a row count within ± 3 rows of being accurate), DP would be achieved. However, the tradeoff to adding this statistical noise is accuracy itself – by preserving differential privacy, we would potentially lose some level of accuracy. In short, DP works to blur the decision boundaries of the ANN in order to make it harder to reverse engineer how those decision boundaries were created. The central question within this subfield, then, is the best mechanism to quickly achieve DP without losing significant accuracy.

Since the original work by (Shokri and Shmatikov 2015), countless others have followed not only within DP, but also regarding other privacy-preserving techniques.

4.5.2 Other Privacy Techniques: Forming a Taxonomy

Homomorphic Encryption

Like DP, homomorphic encryption was first conceived before ANNs became a popular technique (Rivest, Adleman and Dertouzos 1978). However, soon after (Shokri and Shmatikov 2015) first brought privacy issues generally to light regarding ANNs, some scholars began to attempt to apply homomorphic encryption to the problem of

privacy-preservation; the idea of homomorphic encryption is simply to “encrypt data such that certain operations can be performed on it without decrypting it first.” (Dowlin, et al. 2016). Since the values don’t need to be decrypted to function, then, there is no chance of private information leaking out.

However, while fully homomorphic encryption schemes have been developed (Gentry 2009), one of the key criticisms is that they can significantly slow down the processing time of an ANN. Thus, the key question that modern homomorphic encryption has sought to answer is whether it can be done computationally quickly and be universally applied to any kind of data.

Federated Learning

Federated learning (FL) is a technique created by research scientists at Google in 2017 (McMahan and Ramage 2017). Whereas DP adds randomized noise to query results and homomorphic encryption allows the training data to always remain encrypted, federated learning attempts to create privacy through the complete elimination of centralized training itself. The idea is simple enough: eliminate the need for any kind of centralized database of information to conduct machine learning. Instead, have many client machines conduct ML training locally, then transmit the results of that training to a centralized server. The example they use is with smartphones: individual smartphones would be able to run a relatively small amount of ML training on the device, then transmit the results of that training to the centralized server. However, the training data that each smartphone uses is never sent to the

centralized server. Thus, the centralized server isn't a privacy risk if it never has the private data to begin with.

4.6 Democratic Legitimacy

Democratic legitimacy is perhaps the most unique research thread here in that until about 2018, it had rarely been a focus of any ANN scholarship or even more general machine learning scholarship. Given the centrality of democratic legitimacy to public agencies, this will be the most extensive section of the literature review.

Beyond analyzing the foundational works in the field of democratic legitimacy, this section of the literature will review specifically which *activities* within democratic legitimacy are most relevant to the topic of ANNs. This is an important distinction, because not all parts of democratic legitimacy are clearly relevant to ANN development. For example, democratic elections themselves are obviously of great importance to democratic legitimacy, but machine learning systems (at least at the present time) have little association with such activities.

Each of the preceding threads to this section vary in their balance of assessment between quantitative and qualitative – some are purely quantitative, whereas others are a mix of quantitative and qualitative analysis. However, democratic legitimacy is unique in that assessing “more” or “less” is an almost entirely qualitative activity; there is no democratic legitimacy statistic that can be calculated for a public agency. The purpose of this literature review section, then, is to define what activities are involved in

achieving democratic legitimacy and how relevant those activities are to machine learning and especially ANNs.

However, there is no current, unified literature at the intersection of democratic legitimacy and ANNs. Rather, there are two highly relevant sub-fields which I work to synthesize in this section: traditional (i.e. not related to algorithms or machine learning) democratic legitimacy literature, and the emerging area of algorithmic governance literature. The latter also includes so-called “Fairness, Accountability, and Transparency” literature, also known as FAT. While algorithmic governance literature should (as it grows and expands) eventually make the need for including more traditional democratic legitimacy literature obsolete, we are not yet at this point. Thus, I first dive into algorithmic governance literature to find which activities scholars in the field deem the most important to achieve democratic legitimacy. Then, I supplement those activities with those found in traditional democratic legitimacy literature that are relevant to machine learning.

4.6.1 Foundational Works

The concept of democratic legitimacy itself (ignoring machine learning) in modern scholarship began in the late 1960s/early 1970s (Kriesi 2013, 609). At that time, the predominant worry among scholars was that expanding expectations of the State from people would eventually cause democracies to falter. By the end of the Cold War, however, it appeared as though those worries were wrong, and the field of study went

largely ignored. However, social and economic problems since the 2009 Great Recession have caused a resurgence in literature focusing on the issue

4.6.2 Algorithmic Governance

Although the term ‘algorithmic governance’ was first coined as a concept in 2006, its current definition and usage is more closely aligned with machine learning in government. (Danaher, et al. 2017, 1-2) define *algorithmic governance* as when public agencies (or corporate entities) “...outsource decision-making authority to algorithm-based decision-making systems” which may even be “...able to learn and adapt to any decision-making situation without the need for human input or control.” Algorithmic governance is intricately related to democratic legitimacy because of the need to have human beings in control (or at least “in-the-loop”) of automated decision-making (Koulu 2019, 9-11).

Within the concept of algorithmic governance (and thus democratic legitimacy) stand three primary concepts: fairness, accountability, and transparency (often just referred to as FAT). However, while there is an entire sub-field of “FAT/ML” literature, I only focus on FAT/ML scholarship specifically oriented towards algorithmic governance. Of note, fairness within algorithmic governance literature (and FAT/ML) can mean algorithmic fairness, non-algorithmic fairness, or (most often) both simultaneously.

Challenges Facing Algorithmic Governance

There are several challenges facing algorithmic governance. In particular, (Stoyanovich 2019) argues that the “[l]ack of transparency and accountability threatens

the democratic process itself.” To alleviate this lack of transparency and accountability, the author presents a “Data Transparency Framework” based on a case study of New York City’s efforts to tackle the topic.

Most scholarship about non-algorithmic fairness, transparency, and accountability in algorithmic governance also seem to agree that trade secrets are part of the problem: when source code is kept secret, however necessary for profitability, the public suffers from its inability to determine the effectiveness of a machine learning tool (Katyal 2019). However, agreement on the problem does not imply agreement on the solution (Redden 2018).

Machine Learning in Government: Differing Viewpoints

While there is no disagreement in recent scholarship that non-algorithmic fairness, accountability, and transparency are *important* to the idea of algorithmic governance, there is a wide spectrum of viewpoints as to whether or not these standards are reasonably achievable, and in turn whether or not machine learning has reached a point where its benefits outweigh its costs in public agencies.

On one end of the spectrum is (Coglianese and Lehr 2019). The authors discuss whether machine learning algorithms can meet a sufficiently high standard of transparency (and implicitly accountability and non-algorithmic fairness) to achieve democratic legitimacy. They conclude that new technical advances will allow for enough transparency to meet the standards of democratic legitimacy. Even though such transparency may not exist in full today, the authors assert that “[i]n the future, a

government that makes use of so-called black-box algorithms need not be a black-box government. With responsible practices, government officials can take advantage of machine learning's predictive prowess while remaining faithful to principles of open government.” (Coglianese and Lehr 2019). In short, they argue that while there are problems at the present, these problems are not crippling and should not stop us from learning more about how to increase transparency in public agency usage of algorithms.

Other scholars take a more cautious view. While (Coglianese and Lehr 2019) accept that issues of bias exist, the authors are firmly of the belief that technology will allow us to solve these problems and that the benefits clearly outweigh the costs. In contrast, (Brkan 2019) provides a somewhat different view. Their research specifically focuses on the so-called ‘right of explanation’ found within Europe’s new GDPR laws. They conclude that “...if the algorithm used for decision-making is a neural network, prone to very fast machine learning, it will be close to impossible to explain the reasons behind its decision.” (Brkan 2019, 120-121). While accepting that the future may allow for greater transparency with machine learning systems, Brkan argues that without such transparency there cannot be true legitimacy. Thus, Brkan hopes to wait until we have more explanations behind neural networks before beginning to use them in public agencies.

However, some scholarship takes a much darker interpretation of machine learning in the age of algorithmic governance. One exemplar of this train of thought is

(Valentine 2019), who argues that much of the machine learning used by public agencies today amounts to “social control mechanisms to contain and criminalize marginalized populations.” The author argues that rather than any new revolution in efficiency or effectiveness, current usage of machine learning (and other algorithms) in government agencies often amounts to simply re-codifying historical patterns of discrimination into a technological redlining that “reinforces oppressive social relationships.” Rather than expressing the general optimism of Coglianese or the waiting and caution of Brkan, Valentine asserts that almost any usage of machine learning in public agencies for predictive or policing purposes is *inherently* unjust and should be fought against in courts and through activism.

4.6.3 The Activities of Democratic Legitimacy

While there are scholarly debates as to the intersection of machine learning and algorithmic governance, that may not make up the entirety of what is encompassed within the broader idea of democratic legitimacy. This section will identify the specific *activities* that traditional democratic legitimacy literature can add. The algorithmic governance literature provides us with FAT, but what else is there?

These activities will be extracted from three pieces of literature (discussed below). Any extracted activities must also meet three criteria: (1) the activity is pertinent to democratic legitimacy *as it relates to public agencies*, (2) the activity must be *procedural* in orientation, rather than based on a particular subjective policy

outcome, and (3) the activity must not be entirely covered in a previous research thread.

The first criteria should be self-evident – obviously only those aspects of democratic legitimacy that relate to public agencies (and thus this dissertation) are being considered. This is not to say that other aspects are less important, but rather that their consideration is beyond the scope of this study. For example, democratic legitimacy activities related to passing legislation are not included because the focus here is not on legislative activities. The second criteria is necessary because this study is not meant to advocate for particular policies and indeed explicitly avoids adding principles meant to achieve certain policy end goals. Finally, the last criteria is needed because there will inevitably be some element of overlap between democratic legitimacy and one or more of the existing research threads above.

Kriesi's Democratic Legitimacy Typology

(Kriesi 2013, 617) provides an excellent conceptualization of democratic legitimacy in a matrix model. The author divides conceptions of democratic legitimacy into procedural vs. long term results legitimacy and into input vs. output legitimacy:

Table 3 - Democratic Legitimacy in Input vs. Output

Normative basis	Input Legitimacy	Output Legitimacy
Yes	Procedural legitimacy I: satisfaction with the quality of representative democracy	Procedural legitimacy II: satisfaction with the quality of governance

	(responsiveness and accountability)	(rule of law, impartiality, fairness)
No	Partisan legitimacy: satisfaction with electoral outcome	Outcome legitimacy: satisfaction with policy performance

Among the four types of democratic legitimacy-making activities, all activities covered within **Procedural legitimacy I** and **Procedural legitimacy II** meet all three criteria. However, in the case of fairness here, it's again important to make the distinction between *algorithmic fairness* (which is what the literature review section on fairness covered) and *non-algorithmic fairness*, which is being covered here.

Partisan legitimacy and **outcome legitimacy** fail to meet the second criteria for the same reason: they are subjective based upon what one defines as a “good” outcome. Additionally, partisan legitimacy fails the first criteria since it does not directly touch on public agencies. That is, while a democratic election may cause a public agency to produce a new ANN (or stop producing a new ANN), the standards for whether or not the ANN is well-designed and effectively implemented and managed do not change.

What remains are two activities within *Procedural Legitimacy I* (**responsiveness** and **accountability**) and three activities within *Procedural Legitimacy II* (**rule of law**, **impartiality**, and **non-algorithmic fairness**). Let us put a pin in these activities until the other two key pieces of literature are reviewed.

The RESuME Project

The second study chosen was the *Resources on the European socio-economic model* (RESuME) project (Chiocchetti 2017). Their study reviews five general categories related to democratic legitimacy, each with one or more activities within. The **bolded** activities are those which I argue meet all three specified criteria above (if only some sub-bullets qualify, only the sub-bullets are bolded):

- I. Electoral Authorization
 - a. Universality of voting
 - b. openness and fairness of political competition
 - c. integrity of electoral procedures
 - d. level of participation
 - e. characteristics of electoral system
- II. Direct Citizen Participation
 - a. Referendums
 - b. public consultations**
 - c. access to elected representatives and public officials
 - d. internal party democracy
- III. **Deliberation** (i.e. informed and reasoned agreement between different parties)
- IV. Substantive Representation - the preferences and concerns of citizens in the political system are being met through a variety of mechanisms, such as:
 - a. Similarity
 - b. Delegation
 - c. Accountability**
 - d. Responsiveness**
- V. Constitutional Protections
 - a. Checks and balances
 - b. Subjective rights**
 - c. Procedural protections aimed at protecting the individual from the state**
 - d. Procedural protections aimed at protecting minorities from majorities**

Below are the explanations for why each democratic legitimacy activity noted in the RESuME project was either accepted or rejected.

Table 4 - Summary of Activities: Acceptance vs. Rejection

Activity	Rejection vs. Acceptance	Explanation
Electoral Authorization (all sub-activities)		In general, machine learning systems are not associated (at present) with voting activities or access to voting. They are thus beyond the scope of this study.
Direct Citizen Participation (a, c, d)		Referendums, access to elected officials, and internal party democracy are all activities related to the legislative branch and the Executive Office of the President rather than public agencies.
Direct Citizen Participation – Public Consultations		Unlike the other three activities housed inside Direct Citizen Participation, public consultations are needed not only when passing legislation (which is beyond the scope of this study) but also when public agencies are interpreting and implementing those laws.
Deliberation		This category is similar to public consultations, except it might include other key stakeholders besides the public at large.
Substantive Representation (a, b)		Once again, similarity and delegation are generally relevant to the legislative branch and legislation rather than public agency implementation.
Substantive Representation - Accountability		Accountability is a key feature of any public agency, and is repeatedly mentioned in any literature related to algorithmic governance.
Substantive Representation – Responsiveness		While related to accountability, they are not precisely the same concept. A public agency can have fast responsiveness, but those “responses” can themselves contain little information relevant to maintaining accountability. Likewise, a public agency can have strong accountability but be poor at

		conveying that information quickly and effectively to the public.
<i>Constitutional Protections – Checks & Balances</i>		While checks and balances are of course vital in the US, it is also predicated on focusing on the balance between all three branches of government. As this study focuses exclusively on public agencies, and not on how the branches may conflict with one another, it is not pertinent.
<i>Constitutional Protections – (b, c, d)</i>		The three remaining activities under Constitutional Protections are all clearly relevant: public agencies have a responsibility to protect subjective rights, minority rights against the majority, and individual rights against government encroachment.

AI Now's Algorithmic Impact Assessment

If there were to be a foundational work at the *intersection* of democratic legitimacy and machine learning, it would be the AI Now Institute's Algorithmic Impact Assessment (AIA), which provides guidance to public agencies on how to manage what they refer to as “automated decision making” systems⁹ (Reisman, et al. 2018). While their analysis covers more than just ANNs or even machine learning, it is still the best candidate for being the true foundational work at the intersection of democratic legitimacy and machine learning.

Even though the authors never explicitly use the term “democratic legitimacy”, they specify in their Executive Summary that “[t]he turn to automated decision-making

⁹ AI Now introduced their first Algorithmic Impact Assessment in 2016, and then produced successive iterations in 2017 and 2018.

and predictive systems must not prevent agencies from fulfilling their responsibility to protect basic democratic values, such as fairness, justice, and due process, and to guard against threats like illegal discrimination or deprivation of rights.” (Reisman, et al. 2018, 5). Because of this, and because their work specifically considers algorithms in the context of democratic governance, I believe that it is a critical component to include.

To arrive at their conclusions, the authors conducted a comparative study of other impact assessment frameworks, including from “environmental protection, data protection, privacy, and human rights policy,” and apply the principles of those frameworks to their own AIA (Reisman, et al. 2018, 7). The AIA consists of five phases: pre-acquisition review, initial disclosure requirements, comment period, due process challenge period, and AIA renewal.

First, the pre-acquisition phase “allows the agency and the public to identify concerns that may need to be negotiated or otherwise addressed before a contract is signed.” (Reisman, et al. 2018, 8). The idea behind this first phase is to stop the usage of an automated decision-making system before a substantial investment has been made.

Next, the initial disclosure requirements phase deals with the final outputs from the preceding phase. These include:

- Publishing the agency’s definition of an automated decision system
- Disclosing details of the system, such as purpose, reach, internal use policies, and implementation timeline
- Assessing the system internally for inaccuracy, bias, and harms, as well as establishing ways to address those impacts.

- Proposing a plan for providing access to researchers outside the agency who seek to review the system once it is deployed.

Third, the comment period phase is just as it sounds – the public should be provided with time to provide feedback on the system, as well as establish any concerns. Fourth, the due process challenge phase is a more adversarial version of the comment phase – if concerns were not mitigated in the preceding phase, this phase would set aside time for challenges to an agency’s oversight body or even a court of law. The final phase is renewal, which states that the agency should be required to repeat the previous four stages of the AIA “on a regular schedule,” with the suggestion that every two years is a reasonable timeline (Reisman, et al. 2018, 10).

Aligning AIA Key Elements and Policy Goals

With these chronological stages in mind, the AIA provides five key elements (hereafter referred to as E-1 through E-5) for an algorithmic assessment, plus four key policy goals (hereafter referred to as P-1 through P-4) that any public agency using an “automated decision system” should strive towards (Reisman, et al. 2018, 4-5). They are presented below for ease of later comparison:

E-1. Agencies should conduct a self-assessment of existing and proposed automated decision systems, evaluating potential impacts on fairness, justice, bias, or other concerns across affected communities.

E-2. Agencies should develop meaningful external researcher review processes to discover, measure, or track impacts over time;

E-3. Agencies should provide notice to the public disclosing their definition of “automated decision system,” existing and proposed systems, and any related self-assessments and researcher review processes before the system has been acquired;

E-4. Agencies should solicit public comments to clarify concerns and answer outstanding questions; and

E-5. Governments should provide enhanced due process mechanisms for affected individuals or communities to challenge inadequate assessments or unfair, biased, or otherwise harmful system uses that agencies have failed to mitigate or correct.

P-1. Respect the public's right to know which systems impact their lives by publicly listing and describing automated decision systems that significantly affect individuals and communities;

P-2. Increase public agencies' internal expertise and capacity to evaluate the systems they build or procure, so they can anticipate issues that might raise concerns, such as disparate impacts or due process violations;

P-3. Ensure greater accountability of automated decision systems by providing a meaningful and ongoing opportunity for external researchers to review, audit, and assess these systems using methods that allow them to identify and detect problems; and

P-4. Ensure that the public has a meaningful opportunity to respond to and, if necessary, dispute the use of a given system or an agency's approach to algorithmic accountability

4.6.4 Synthesizing Democratic Legitimacy Activities

In this section, I take the activities identified in the algorithmic governance and traditional democratic legitimacy literature and merge those which are functionally identical or split those covering too big a subject area. I then place the results in a matrix summary table for the final activities list of democratic legitimacy.

Initially, we have the following democratic legitimacy-seeking activities from the democratic legitimacy literature review above:

- Public consultations
- Non-algorithmic fairness
- Transparency
- Deliberation

- Accountability
- Responsiveness
- Subjective rights
- Procedural protections (individual against the state)
- Procedural protections (minority against majority)
- Impartiality
- Autonomy
- Interpretability
- AIA's E-1 through E-5
- AIA's P-1 through P-4

However, there is still substantial overlap in these activities, and they can be refined into a much smaller group.

Deliberation vs. Public Consultations

These are both similar concepts, with the core difference between who is being consulted: public consultation requires the consultation of the population at large, whereas deliberation can involve just about any group with an interest in a given policy. Since deliberation can effectively include public consultations within it, public consultation will be considered an element of deliberation.

Procedural Protections

Procedural protections for the individual against the state and for minority groups against the majority can be more succinctly be stated as substantive due process rights under the 5th and 14th Amendments (Broderick 2009). The “famous footnote” of the *United States v. Carolene Products Co (1938)* case made clear that the legislative branch (and the federal agencies which implemented its legislation) would be under “strict scrutiny” for actions which fell under the following categories:

- Activities which violated the Bill of Rights

- Activities which violated fundamental political processes, such as voting and free speech
- Activities which were prejudicial against “discrete and insular minorities”

Additionally, we can add subjective rights (defined as those referenced in the UN’s 1948 Universal Declaration of Human Rights) to this list (Peters 2011). Moving forward, procedural protections for individuals and minority groups plus subjective rights will be jointly referenced as substantive due process rights, *or SDPR*.

Responsiveness vs. Accountability

Both (Chiocchetti 2017) and (Kriesi 2013) place responsiveness and accountability together (in their substantive representation and procedural legitimacy I sections respectively), and this makes sense. While the two terms aren’t identical, they are intrinsically related to one another – being responsive is necessary to being accountable and being accountable is necessary to being responsive. For the purposes of this study, responsiveness will be considered an element of accountability.

Transparency vs. Accountability

While there is an element of intersection between transparency and accountability, they are not equivalent. Rather, transparency is a necessary but not sufficient prerequisite for accountability (Koene, et al. 2019, 1). In other words, transparency can exist without accountability, but accountability cannot exist without some element of transparency. However, transparency still has value as a separate activity – while achieving transparency without accountability is obviously not preferable, such a result is certainly preferable for a public agency than achieving neither. Therefore, I will leave them as separate activities.

Interpretability vs. Explainability

For the purposes of this study, explainability refers to what we can explain or understand of the model itself (i.e. explanatory power), whereas interpretability is more closely related to how a public agency provides that explanation to the public (or at least to those groups that require explainability). An ML system may have perfect explanations provided, but a public agency can still fail to make those explanations meaningful to the public. Likewise, a public agency may do its utmost to effectively share as much as possible about the ML system that it uses. However, if it simply doesn't have much explainability to begin with, such interpretability isn't as useful. In short, there is an overlapping and complementary relationship between the two, but they are nevertheless distinct: explainability can be covered algorithmically in the literature review above, whereas interpretability more closely relates to the public agency's activities rather than work done on or with the algorithm/data.

Integrating AIA Elements & Policy Goals

Given the AIA's specific focus is on algorithms, it stands to reason that their democratic legitimacy-seeking activities are more specified and focused than the activities for the other two studies reviewed. Because of this, some of them can be fit squarely into one or more of the previously defined activities. Specifically:

Table 5 - Linking Legitimacy Activities to the AIA

Current Legitimacy Activity	Relevant AIA Element/Policy Goal
Deliberation	E-4, P-3, P-4

Accountability	E-2, P-2, P-3, P-4
(Protection of) Substantive Due Process Rights	E-1 (partial), E-5, P-4
Transparency	E-3, P-1
Non-algorithmic Fairness	E-1, E-2, E-5, P-2
Interpretability	E-3, P-1, P-3

However, the AIA study also has two activities that are not represented in the preceding two pieces of literature. In particular, this refers to algorithmic maintainability and human autonomy.

Algorithmic Maintainability

In addition to the activities specified from the primary democratic legitimacy literature reviews, I argue that there is one additional activity required in the context of ANNs and machine learning generally: algorithmic maintainability. Both E-2 and P-3 within AIA's study help describe algorithmic maintainability. From them, I define algorithmic maintainability as *"the process through which a machine learning system is reassessed to ensure that it continues to meet or exceed previously approved standards of performance, both in terms of accuracy and other relevant assessment metrics."*

With human-made decisions, there are natural processes of maintainability: for example, laws or regulations can be changed over time as society changes, and new leaders are routinely elected. However, ML systems lack this kind of maintainability. If ANNs are going to be used in potentially important decision-making processes, then it is

important for them to have the need for their maintainability spelled out for them to remain democratically legitimate.

At the same time, ML systems shouldn't necessarily be treated as generic software programs – updating them (i.e. adding new data to train from) simply for the sake of updating isn't necessarily a net positive – those updates could potentially lower accuracy or even have lower quality data. Then there are subjective questions as to how, precisely, to do the update – should old data be thrown out, should the new data simply be added on top of the old data?

Human Autonomy

Along with algorithmic maintainability, human autonomy is another unique concept presented in the AIA study, a concept which is specific to the issues created with machine learning systems. Human autonomy in relation to ML systems is defined as an individual's capacity for self-determination or self-governance. More specifically, P-1 and P-4 within the AIA study both reference how ML systems might harm an individual's capability for self-determination.

4.6.5 Key Democratic Legitimacy-Inducing Activities Defined

With these key activities refined and integrated together, the final list of eight key democratic legitimacy activities as related to ML systems is below:

Table 6 - Definitions of Democratic Legitimacy Activities

Democratic Legitimacy Activity	Definition
Deliberation (i.e. deliberative democracy)	"...political decisions should be the product of fair and reasonable discussion and debate among citizens." (Eagan 2013)
Accountability	"A set of mechanisms, practices and attributes that sum to a governance structure which involves committing to legal and ethical obligations, policies, procedures and mechanism, explaining and demonstrating ethical implementation to internal and external stakeholders and remedying any failure to act properly" (Koene, et al. 2019, 4)
Substantive Due Process Rights (i.e. SDPR)	The public agency, to the best of its ability, ensures the substantive due process rights of individuals who are assessed with its machine learning system are not violated. These rights specifically include: - Activities which violate the Bill of Rights - Activities which violate fundamental political processes, such as voting and free speech - Activities which are prejudicial against "discrete and insular minorities" - Activities which violate the rights specified in the UN Universal Declaration of Human Rights
Algorithmic Maintainability	The process through which a machine learning system is reassessed to ensure that it continues to meet or exceed previously approved standards of performance, both in terms of accuracy and other relevant assessment standards.
Transparency	"Depending on the type and use of an algorithmic decision system, the desire for algorithmic transparency may refer to one, or more of the following aspects: code, logic, model, goals (e.g. optimisation targets), decision variables, or some other aspect that is considered to provide insight into the way the algorithm performs. Algorithmic system transparency can be global, seeking insight into the system behaviour for any kind of input, or local, seeking to explain a specific input - output relationship." (Koene, et al. 2019, 4)

(Human) Autonomy	“Autonomy is an individual’s capacity for self-determination or self-governance.” (Dryden n.d.)
(Non-algorithmic) Fairness	In contrast to optimizing for fairness in an algorithm, this conception of fairness deals with how a public agency itself implements fairness at an organizational level rather than at an algorithmic level. This includes concepts such as ensuring that the data used to train is fair to different sub-groups, as well as ensuring that all agency employees who utilize a machine learning system are well trained.
Interpretability	Interpretability is one of the least well-defined terms in machine learning. In a thorough review of existing interpretability literature, (Lipton 2016) argues that the objectives of interpretability include trust, causality, transferability, informativeness, and fair and ethical decision-making.

4.6.6 Separate Research Thread vs. Democratic Legitimacy Activity

Among the democratic legitimacy-seeking activities noted above, four stand out as being potentially viable candidates for their own research thread: algorithmic maintainability, transparency, non-algorithmic fairness, and accountability. However, each was rejected for the same reason: while they are intricately related to ANNs (and machine learning more broadly, to varying degrees), they are not *primarily technical properties* to be optimized. That is, they will not be solved (at least primarily) through enhancing the algorithm, the model, or the data, but rather through the structure and procedures of the public agency itself. This contrasts with accuracy, robustness, privacy,

explainability, and (algorithmic) fairness, each of which are primarily technical in how they are optimized.

4.7 Excluded Possible Threads

While each of the six threads listed above are included in my analytical framework, there are several other candidates that were considered as well, either as separate research threads or as activities housed within democratic legitimacy.

4.7.1 Ethical AI

Ethical AI literature focuses on how AI should be implemented in society more broadly. While this topic is not included in my literature review directly, a meta-analysis of different ethical AI frameworks is included in my formal methodology below (Stage Two).

One of the most important works in this field is from (Mittelstadt, et al. 2016), who defined six distinct ethical concerns for algorithms. While their work is for algorithms generally, it's nevertheless applicable to neural networks as well. These six ethical concerns include: inconclusive evidence (there will never be 100% accuracy), inscrutable evidence (hidden connection between data and conclusion), misguided evidence (how representative the data is of reality), unfair outcomes (standards of fairness and bias), transformative effects (the oft-hidden changes in how people conceptualize their world due to algorithms), and traceability (determining who or what is responsible for adverse decisions). Many of these ethical concerns can be mapped to

previous areas of neural network scholarship. Because a study of Ethical AI frameworks is included in the formal methodology as an archival review, I found it unnecessary to include it as a section of the literature review as well.

4.7.2 Law & Regulation

There have already been a variety of policy and legal scholars who have weighed in on the implications of implementing ML systems in society, both in terms of how to regulate emerging technologies (Bonnín-Roca, et al. 2017) (Price II 2017) and whether or not various uses of machine learning are even legal for a public agency to use (Coglianese and Lehr 2017).

However, the reason why this thread was not included is because it is beyond the scope of this study. This dissertation is not focused on answering questions related to the proper way to regulate or deal with the legality of ML systems. Additionally, questions of legality and regulation move away from the procedural side and looks at the end result. Rather, my hope is that researchers interested in these fields can draw from my analytical framework to develop effective laws and regulations.

4.7.3 Behavioral Psychology

Behavioral psychology is another thread of research where there is unfortunately little in the way of published research related to machine learning systems, particularly those used in the context of public agencies. Indeed, even many recent AI ethical frameworks have failed to delve into this field very deeply. Because of

the lack of significant peer reviewed scholarly publications available at the intersection of ANNs and behavioral psychology, it makes this area untenable as a separate research thread in the literature review. In short, this thread was not included because it still needs more basic research when applied to ANNs, not to mention being outside the field of public policy. Were this a dissertation in the field of behavioral psychology, it would be quite a different story. Indeed, once more research has been conducted, I believe it would be an excellent and highly useful thread to this literature review (or potentially added within democratic legitimacy).

However, despite my inability to find much peer reviewed scholarly literature at this intersection, there is still a smattering of news articles and blogs on the subject which allow for a basic review of concepts. There is also some scholarly literature focusing on the intersection of automated systems more generally with behavioral psychology.

Within behavioral psychology, then, the primary issue relevant to artificial neural networks is that of *cognitive bias*. However, this should not be confused with the previous section discussing bias as it relates to algorithmic fairness. Whereas that section focuses on bias from the neural network side of the equation (such as an ANN being more likely to misclassify input data from a minority group), cognitive bias looks at the human side of bias.

According to Cami Russo, author of Psychology Today's *The Future Brain: The Intersection of AI and Human Intelligence* series, "[h]uman cognitive bias influences AI through data, algorithms and interaction." She further notes that "[t]he size, structure, collection methodology, and sources of data impact machine learning. Machine learning is dependent on the quality of learning data sets." (Russo 2018). These three areas of influence are important to focus on: data, algorithms, and interaction.

First, data can be subject to cognitive bias because even if an artificial neural network determines for itself which input parameters are important, it is still a human being who defines the overall list of parameters to choose from in the first place. How a computer programmer conceives of what parameters might be important, even if the ANN itself determines which subset of parameters are the most predictive, can fundamentally shape the result.

Second, algorithms can be subject to cognitive bias because of the sheer variety of hyperparameter choices that a human being must select from. Choices include type of layer (RNN, LSTM, CNN, fully-connected, etc.), number of layers, number of neurons per layer, activation function, and more. The cognitive assumptions a programmer has about which algorithms will perform best at which task will deeply shape the final ANN. Indeed, this problem of hyperparameter selection is well-known in scholarly research, to the point where some scholars have attempted to develop automated machine learning solutions for hyperparameter selection itself (Rodriguez 2018).

Finally, human interaction with an ANN's output can be shaped by cognitive bias. Consider the hypothetical case of an ANN used to assess patient risk at a hospital – the ANN is implemented to predict if a patient is likely to soon need medical assistance based on blood pressure and other relevant medical metrics in real time. It then outputs that likelihood for the presiding doctor as a standard percentage likelihood (i.e. “patient X has a 60% chance of needing medical assistance within the next hour”). Presuming that the system is highly accurate, unbiased, and mitigates all the other problems noted in the threads above, the cognitive bias of human interaction can still play a role. This is because human beings are less than ideal at acting “correctly” based on raw statistics (Rosenblat, Kneese and Boyd 2014). Indeed, the very usage of an automated system *itself* may lead to new human biases, which some refer to as *automation bias* (Skitka 2011). At the same time, other scholarship has asserted that a well-designed artificial neural network may actually be a countervailing force against intrinsic cognitive biases (Andreessen Horowitz 2017).

5 Research Methodology

With the various taxonomies of AI, background information, and a wide literature review in hand, below is my research methodology for conducting my study.

5.1 Introduction

My research methodology is a qualitative, multi-method, iterative approach consisting of archival research, comparative analysis, expert interviews, and peer review which refines and improves my analytical framework. It consists of five distinct, sequential, and interrelated stages:

- (1) evaluate the relationship of the competing research threads noted above to one another;
- (2) extract key actionable principles from existing “ethical AI” frameworks in various fields of study;
- (3) develop a draft analytical framework based on the previous two stages;
- (4) conduct a combination of peer review and expert interview to iteratively improve the framework;
- (5) compare what has been produced at this point against an existing similar framework for public agencies for additional iterative improvements.

The first stage allowed me to take the key elements of my literature review and ask a very simple question of them all: while all of these research threads are important to optimize, particularly in a public policy setting, how do they interact with one another? That is, does optimizing for one cause problems for another? What does current empirical scholarship have to say on the subject (limited though it is at times), and how can I add on to this literature with an analysis of democratic legitimacy as a new thread?

The second stage then looked at those ethical AI frameworks that have been developed thus far from different fields of study: general ethical AI in society, law, public administration, software development, and government grand strategy. Based on several criteria (discussed below), I then selected the most relevant principles from the ethical AI frameworks to utilize in my own analytical framework.

Next, the third stage was the construction of the first draft framework itself – from the conclusions drawn in the first two stages, I generated my initial principles. Some of those principles had a one-to-one relationship with the principles extracted in the previous stages, whereas others were based upon derivations or combinations of multiple concepts. At the conclusion of the third stage, my *first draft analytical framework* was completed (this is located in *Appendix A-1*).

The fourth stage revolved around iteratively improving my analytical framework through a combination of expert interview and peer review. At the conclusion of the

fourth stage, I had my *second draft analytical framework* completed (this is located in *Appendix A-2*).

Finally, the fifth stage encompassed a comparative analysis with the framework produced by (Leslie 2019), the only directly comparable work focused on the same set of problems to this study's analytical framework, for further iterative improvement.

5.2 Stage One: Testing of Competing Research Threads

This stage consisted of an archival review of empirical literature (save for democratic legitimacy, where there isn't necessarily empirical literature) where the relationship between two or more research threads were tested against one another. Specifically, I looked for literature which showed either a positive (i.e. complementary), negative (i.e. in tension), or mixed relationship between two or more research threads when attempting to optimize them. These terms are defined below:

Complementary Relationship: We would expect optimizing one research thread to have a *positive* impact on the other research thread, or that optimizing for both simultaneously would not have a negative impact on either thread compared to them being optimized separately.

Tension Relationship: We would expect optimizing one research thread to be in tension with another if it simultaneously decreases the optimization of a separate research thread, or has a negative relationship.

Mixed Relationship: Sometimes the relationship is positive and sometimes the relationship is negative depending on conditions, or the relationship is entirely neutral, or the relationship cannot be determined.

5.2.1 Defining Optimization

For each of the research threads, the meaning of “optimization” is different. While a given definition of optimization may be best practice today, that does not mean that it will continue to be the best method of determining how optimized a given research thread is in the future. Indeed, for some research threads there is no universally accepted objective standard for what optimizing it even looks like. Additionally, because this is a relatively new field of study within artificial neural networks, the existing literature in this area is not always particularly deep, particularly when looking for direct empirical evidence.

Below are definitions for how I define optimization for each research thread:

Accuracy

Optimizing for accuracy involves assessing either the simple accuracy, the recall/precision/F1 Score, or alternative accuracy replacement measurements in unique sub-fields such as natural language processing (such as BLEU).

Privacy

There are many ways of optimizing for privacy in ANNs. This study considered optimizing for differential privacy primarily (since that is where the highest

concentration of literature exists) but also allowed for studies that looked at secure enclaves or other newer methods if they could be found.

Robustness

This study considered certified robustness standards as optimizing for robustness. As those robustness standards are focused on defending against adversarial examples and data poisoning, that is how robustness' optimization is assessed in this study.

Fairness

Fairness is arguably the thread with the most complicated debate over the proper definition, which in turn changes how it should be optimized. Indeed, many definitions are mathematically incompatible with one another. Rather than simply selecting one definition of fairness as legitimate, the empirical studies comparing the optimization of fairness against other research threads is thin enough that regardless of which standard of algorithmic fairness is chosen, the study will be included here.

Explainability

With explainability, it is particularly difficult to define what optimization looks like because it is inherently less quantitative in nature than the other research threads (except for democratic legitimacy). Indeed, the most "optimized" explanation for a given ANN's output can be different depending on the situation, and determining if an explanation got better or worse is hardly an entirely mathematical exercise. Because of this, it will be mostly left out of this section, except for comparing it against democratic legitimacy.

Democratic Legitimacy

Whereas explainability has both quantitative and qualitative components to it, democratic legitimacy has entirely qualitative components. Math will not directly determine if an ANN created by a public agency has democratic legitimacy (though it may indirectly assist by helping prove issues such as fairness). Because of this, determining “optimization” for democratic legitimacy against other threads will not be empirically based. Rather, we will qualitatively assess the different democratic legitimacy activities defined previously in *Section 4.6.5*. This will be explored in Chapter Six in more depth.

5.3 Stage Two: Ethical AI Framework Meta-Analysis

Unlike the remaining Stages of research, Stage Two does not build directly on top of Stage One. Rather, these two stages were conducted independently of one another with the intent that each shed light onto different kinds of principles, questions, and concepts. Together, they form the baseline first draft of my analytical framework. Whereas the previous stage looked at weighing research threads against one another, this stage analyzed the myriad of “ethical AI frameworks/principles” that have been created by various individuals, groups, and governments. There are dozens of such frameworks that have already been written, and from a wide range of fields. While this section is not intended to be entirely exhaustive of every framework in existence, it should nevertheless cover a wide swath of what presently exists. Indeed, it even covers several other meta-analyses of ethical AI frameworks.

The purpose of this stage was two-fold. First, it determined areas of deep agreement or disagreement between various frameworks. Conflicts may indicate principles without wide areas of scholarly agreement, whereas deep and widespread agreement can indicate concepts which are widely accepted enough to potentially become principles for this framework. Second, analyzing these principles helped to provide guidance in Stage Four during expert interviews.

The remainder of this Stage contains three sections: (a) listing the AI ethics frameworks chosen for this analysis and why they were chosen when others were not, (b) describing the different types of sources for AI ethics frameworks, and (c) the principle criteria that will be used to assess each frameworks' principles for validity and use in this study.

5.3.1 Listing of AI Frameworks Covered

Mostly since 2016, dozens of groups have attempted to develop ethical AI frameworks. One of the most prominent examples was in January 2017, when AI researchers from all over the world met to discuss the implications of artificial intelligence (primarily ANNs and other 'black box' machine learning systems) going into the future. Following the completion of their conference, they codified what they considered to be the 23 most important principles that all AI researchers should follow in their own research (Future of Life Institute 2017). By July 2018, over 1,250 AI researchers (among them many of the leading scholars in the field) had signed onto

those principles. For any scholar looking to analyze the current state of understanding of the impact of ANNs on society, the Future of Life Institute’s 2017 conference principles on AI is an excellent starting point.

Their 23 principles included five dealing with research itself, thirteen dealing with the ethics and values of developing AI, and five assessing the long-term applications and implications of AI. Although some of these principles are too vague to provide any substantive guidance (“There should be constructive and healthy exchange between AI researchers and policy-makers”), others are more meaningful in their direct implications (“Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority”).

The table below includes the frameworks I analyzed. Most of them are actual frameworks themselves, whereas a few are meta-analyses of various previous frameworks.

Table 7 - List of AI Ethics Frameworks

Framework/Paper Name	Citation	Source of Framework
AI at Google: Our Principles & Responsible AI Practices	(Pichai 2018) (Google 2019)	Software Development
The UX of AI	(Lovejoy 2018)	Software Development
AI UX: 7 Principles of Designing Good AI Products	(Pásztor 2018)	Software Development

Montréal Declaration for Responsible Development of Artificial Intelligence	(Abrassart, et al. 2018)	Government
Executive Order on Maintaining American Leadership in Artificial Intelligence	(Trump 2019)	Government
Governance Principles for a New Generation of Artificial Intelligence: Develop Responsible Artificial Intelligence	(Laskai and Webster 2019)	Government
Ethics Guidelines for Trustworthy AI	(High-Level Expert Group on AI 2019)	Government
ALGORITHMIC IMPACT ASSESSMENTS: A PRACTICAL FRAMEWORK FOR PUBLIC AGENCY ACCOUNTABILITY	(Reisman, et al. 2018)	Public Administration & Law
Machine Learning for Public Administration Research, with Application to Organizational Reputation	(Anastasopoulos and Whitford 2019)	Public Administration & Law
Regulating by Robot: Administrative Decision Making in the Machine-Learning Era	(Coglianese and Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era 2017)	Public Administration & Law
AI and Its Impact on Public Administration	(Shrum, et al. 2019)	Public Administration & Law
Asilomar AI Principles	(Future of Life Institute 2017)	Civil Society
TOP 10 PRINCIPLES FOR ETHICAL ARTIFICIAL INTELLIGENCE	(UNI Global Union 2017)	Civil Society
The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems	(Amnesty International 2018)	Civil Society
Universal Guidelines for Artificial Intelligence	(The Public Voice 2018)	Civil Society
The Ethics of AI Ethics: An Evaluation of Guidelines	(Hagendorff 2019)	Meta-framework

Introducing the Principled Artificial Intelligence Project	(Hilligoss and Fjeld 2019)	Meta-framework
--	----------------------------	----------------

5.3.2 Four Existing Sources of Frameworks

In the table above, we can see AI frameworks coming out of four existing sources, as well as “meta-frameworks” which attempt to do something like what this Stage is doing. The meta-frameworks are included to help analyze additional frameworks not directly included in this Stage. The four sources are defined as such:

Civil Society

Civil society frameworks are created by non-governmental non-profit organizations or by groups of scholars in general. Some are peer reviewed, whereas others are not.

Public Administration & Law

These frameworks are generally from legal scholars. They focus on the legal aspects of implementing all kinds of machine learning in society – what laws already exist, what laws might need to be created, where existing case law is headed, and what kind of regulations currently exist which deal with government agencies handling ML systems.

Software Development

These principles are created by software developers. In general, their frameworks are more oriented towards the technical minutia of developing ML systems, as well as how those technical minutia interact with society. They often provide guidance for the software developers themselves seeking to develop and implement ML systems in their own companies.

Government

While these “government” documents sometimes discuss AI in other facets besides ethical principles in society, they are also basic primers for how major world powers see the development of ethical AI, or at least what they are saying publicly about it. While other countries have also produced such AI ethics guidelines, I focus on four governmental entities in particular: the US, Canada, China, and the European Union (EU). Aside from the obvious inclusion of the US, Canada was chosen because of their history in developing ANNs, China was chosen because along with the US they are one of the leaders in AI R&D spending, and the EU was chosen due to their GDPR law and its potential worldwide impact on ANNs and other ML systems. Although this study is focused on US public policy and public agencies, that does not mean that principles noted in these other countries are irrelevant.

5.3.3 Principle Assessment Criteria

With the list of AI ethics frameworks to be considered in hand, I will apply the following criteria to them to filter and extract out those principles deemed most

relevant. A potential principle must meet all five criteria to be considered for inclusion in this framework. Principles on the ‘borderline’ of one or more criteria are discussed further below. The five principle assessment criteria for inclusion are as follows:

Criteria 1: The principle is not just a computer science principle

Principles revolving exclusively around computer science were not selected.

There are already countless guides to selecting the proper number of layers or neurons, selecting the proper optimization algorithm, how many iterations (epochs) to allow an ANN to be trained, how large the batch size should be, etc. What’s more, such principles may only be right today – they are the kinds of principles that lend themselves to change over time. Additionally, they are outside the scope of this framework – this framework is meant to be at the mezzo-level of analysis, not the micro-level. It is not meant to inform computer scientists on the technical specifics of ANN development. Rather, it is meant for public agency managers (and analysts) who may not be as well versed in computer science.

Criteria 2: The principle is relevant to US, domestically-focused public agencies

Some principles may revolve principally around private sector actors, non-governmental bodies, or US public agencies focused abroad. While they may be valid principles for their intended target, they are not relevant for this analytical framework given its scope limitations.

Criteria 3: The principle is not overly generalized, self-evident, or simply inaccurate

Principles that provide little other than generic platitudes or are deemed to be so self-evident that they provide little to no practical utility were not included. This is a

common issue for many AI ethics frameworks. As (Whittlestone, et al. 2019) note in their meta-analysis of ethical AI frameworks, many AI ethical frameworks “...are often too broad and high-level to guide ethics in practice,” such as those that mention that “AI should be used for the common good, should not be used to harm people or undermine their rights, and should respect widely held values such as fairness, privacy, and autonomy.” Therefore, the principle must at least be partially actionable to be included – there should be specific activities that a public agency manager could pursue (or avoid pursuing) that are directly guided by a given principle. For example, there are few who would argue against fairness, good explanations, high accuracy, robustness, and privacy being “good”, or that bias, black boxes, and discrimination are “bad”. But simple statements of “good” and “bad” are insufficient for inclusion.

Criteria 4: The principle does not violate democratic legitimacy

For a public agency, democratic legitimacy is of the utmost importance. This is admittedly an assumption throughout this study, that all public agencies should seek democratic legitimacy. However, I do not believe it to be a poor or improper assumption to make. Nevertheless, potential principles originating from different sources where democratic legitimacy is not necessarily a primary consideration (or simply may not have been considered in such terms) may therefore require special attention before inclusion.

Criteria 5: The principle does not rely on normative assertions of the results of specific policies

As has been noted previously, this framework is designed to be *procedurally normative* – that is, it makes normative assertions about the proper procedure to follow when developing and implementing ANNs and to some extent ML systems more broadly, regardless of what one's intended policy goals are. Indeed, democratic legitimacy itself is viewed through a procedural lens in this framework. Because of this, principles which are aimed at achieving one particular policy goal were excluded. This is not to make a judgment that such policy goals are illegitimate, but rather that they are beyond the scope of this framework.

As an example, this framework is not itself meant to assess whether the police or the FBI *should* be permitted to utilize ANN-based facial recognition software when searching face matches for criminals, or whether they should be permitted to pull in photos from social media and the internet generally (Collins 2019). Thus, a principle which asserted that ANNs should or should not be used by police for facial recognition would not be included in this study's analytical framework. While this issue is undoubtedly a vitally important question of public policy, it is outside the scope of this framework.

Rather, this framework is designed to ensure that if such a policy goal *was* desired, there are clear normative procedures which could mitigate as many negatives as possible during development and implementation. In short, this framework tries to

make no judgment on whether a policy end goal *itself* is normatively good or not. Of course, even if every procedurally normative principle in this framework were to be followed, someone can still argue that the resultant policy is wrong for ethical or moral reasons related to the policy outcome.

5.4 Stage Three: Produce Draft Analytical Framework

With the first two stages complete, the next stage is to produce the first draft analytical framework. The determination of what qualifies as a principle and what principles should be included is inherently a subjective one, regardless of how rigorous and transparent the selection procedures, but I argue that the criteria I have set ensures that this framework will have the greatest possible utility to US public agencies. However, this initial draft is just that – a draft, and an early one. This early draft is located in *Appendix A-1* rather than in Chapter Six to avoid confusion.

5.5 Stage Four: Evaluate, and Improve Analytical Framework

With the preceding three stages complete, I have my *first draft analytical framework* in hand. From there, the task moves on to evaluation and iterative improvement of that framework.

5.5.1 Evaluating the Draft Analytical Framework

Evaluation is of critical importance when developing any kind of framework. However, the challenges of evaluation are potentially tricky in this case – while there is a wide body of literature concerning qualitative evaluation methodologies generally, most

of those methodologies are not designed to *themselves* evaluate a framework. Rather, they generally focus on program or project evaluation. This being the case, I instead developed an evaluation methodology based on RAND Corporation's 2013 evaluation methodology development framework (Guthrie, et al. 2013). In this fourth stage, I analyze how my draft analytical framework would fit into RAND's evaluation methodology framework, consider which methods and tools are best suited to evaluating my analytical framework, and then implement the evaluation methodology it recommends.

RAND's framework itself utilizes a case study analysis and comparison of fourteen previous research evaluation methodologies to build their evaluation methodology development framework. In this section, I outline how my own analytical framework fits into their evaluation methodology.

RAND first determines four central types of characteristics based on what they deem to be the most important in an evaluation methodology: summative vs. formative evaluation, purpose of what is to be evaluated, types of tools used in the evaluation, and in what stage(s) the research should be measured, either quantitatively or qualitatively. Based on these four central characteristics, the authors make separate recommendations for developing an evaluation methodology.

With these characteristics in mind, they then ask the framework's user to answer thirteen questions that further shape the users' evaluation methodology. These

questions include topics such as: purpose, characteristics, context, pitfalls, tools, level of aggregation, and implementation. Below, I provide my conclusions based on what I determined the four central characteristics of my analytical framework are, with particular emphasis given to the choice of tools (i.e. methods).

5.5.2 Key Characteristics to Evaluate

The first central characteristic RAND defines is whether you want a *summative or formative* evaluation methodology. As the names imply, summative evaluations assess what currently exists, while formative evaluations “focus on learning and improvement rather than assessing the current status” (Guthrie, et al. 2013, 5). My analytical framework was incomplete after the first three stages are finished. Therefore, a formative evaluation methodology is obviously fitting.

The second central characteristic the authors define is based the purpose of the evaluation itself. They define four generalized purposes of evaluation: advocacy (making the case for the program being evaluated), accountability (determining whether funding for a given project was used effectively), allocation (determining how much funding to allocate for a given program), and analysis, which they define as “to understand how and why research is effective and how it can be better supported, feeding into research strategy and decisionmaking by providing a stronger evidence base” (Guthrie, et al. 2013, 6)

I argue that while none of the four are perfect fits, *analysis* is clearly the best fit for my evaluation methodology. This is because the other three types are primarily concerned with project management evaluations, whereas the analysis typology is more closely aimed at research itself. Also, an *analytical* framework should naturally be concerned with analysis first and foremost. However, it is admittedly not a perfect fit, but research and development is inherently part of what a public agency would require when implementing an ANN or other ML system. At this point, based on RAND's methodology, my evaluation methodology would be a *formative analysis*.

The third central characteristic defined by Gurthrie et al is at what *stage* the measurement itself should take place. They point to five possible stages of measurement:

- Input measures, which capture the resources consumed for an intervention to take place
- Output measures, which accounts for the goods and/or services directly produced as a result of an intervention
- Process measures, which capture what occurs between input and output
- Outcome measures, which reflect the initial impact of an intervention
- Impact measures, which reflect the long-term impact of an intervention.

However, not all types of measurement are relevant for all evaluation methodologies. For the purposes of my analytical framework, I argue that *input measures* and *impact*

measures can be immediately discounted. The former can be discounted because the input is minimal – the resources required to utilize my analytical framework (as compared to not using it) is not directly financially significant. This should not be confused with the costs associated with developing an ANN or ML system, which could be significant. Likewise, the latter can be discounted because there hasn't been enough time for ANNs and black box ML systems in public agencies to even conduct such an extended analysis.

This leaves us with measurement at three possible stages: process, output, and outcome. However, among these three outcome is not be as viable as the other two since I am not actually implementing my framework in a real-world public policy situation – I do not control the levers of government, and thus *outcome* is particularly difficult to assess since I won't myself be able to place a neural network created through my analytic framework in a real-world situation as a part of a government agency. That stage of evaluation will be left to those in public administration, and should prove a fruitful avenue of future research.

Thus, measured my framework at two stages: process and output. In my case, the process includes the overall quality of the methodology for creating this analytical framework, and the output includes the final principles themselves.

Finally, (Guthrie, et al. 2013, 9) unsurprisingly consider the choice of tool(s) to be a particularly important characteristic in an evaluation methodology. Their framework

includes “Group 1 tools,” which includes case studies, documentary review, site visits, and peer review, and “Group 2 tools,” which includes bibliometrics, economic analysis, interviews, and data mining. Group 1 tools are generally “formative, flexible and able to deal with crossdisciplinary and multi-disciplinary assessment,” whereas Group 2 tools are generally “scalable, quantitative, transparent, comparable, free from judgement and suitable for high frequency, longitudinal use.” (Guthrie, et al. 2013, 9).

Considering the formative nature of my evaluation methodology, as well as the cross-disciplinary nature of my research generally, my tools should come from Group 1. While expert interviews is arguably included in Group 2, from context it appears that Guthrie et al. are referring to more *en masse* interviews with a broader population, rather than in-depth interviews with a select group of experts.

There are several reasons for this selection. First, most quantitative metrics (i.e. Group 2) are less relevant or effective in evaluating my analytical framework. This is because there is no straightforward quantitative metric, such as accuracy, whose improvement would be strongly correlated with my analytical framework “improving.” There is also no measurement of profitability or any easy way to quantitatively measure “success”, either.

Second, many other qualitative methods for evaluation fall short in evaluating my analytical framework. For example, *site visits* to current sites of neural networks being used in public agencies would be unlikely to be effective or even plausible given

the general restrictions on divulging information about such systems. Indeed, previous studies even attempted to obtain such information (on either ANNs or machine learning generally) with FOIA requests, yet were generally unable to pierce the lack of transparency that exists today (Brauneis and Goodman 2018).

Finally, peer review and expert interviews provide a way to “escape the bubble” of theoretical research and ensure the relevancy of my framework outside my own research. This is particularly important for a public policy dissertation with a strong element of computer science. Peer review allowed other scholars to assess and comment on my findings prior to publication, and expert interviews allowed me to delve deeper into what I gather from my peer reviews. Additionally, a comparative analysis allowed me to compare my own research and ATI’s, with each framework having been conceived of and developed entirely independent of the other.

5.5.3 Evaluation Methods Explained

Given the relatively restricted group of individuals to choose from (see below), the lack of available financial compensation for interviewees’ time, and the comparatively large request *of* their time, I considered a minimum of five interviewees/peer reviewers to be acceptable.

Participant Selection Criteria

My participants all needed to meet the following minimum requirements:

- The participants were from differing fields (or if from the same field, then covering an entirely different subset of that field), but each would need a history of interdisciplinary peer-reviewed publications involving the impact of artificial neural networks (or machine learning generally) on society.
- The participants could either be computer scientists with a deep interest in the social sciences, or social scientists/legal scholars with a deep interest in machine learning.
- All interviewees were working on issues relevant to the United States, given this study's domestic focus.
- All interviewees must either have had a PhD or JD already, or be a PhD candidate with a history of relevant first-author peer reviewed publications
- All interviewees must be 21 years of age or older
- All interviewees must sign an authorization from the George Mason University Institutional Review Board asserting that their participation is voluntary and that they will be recorded

Structure of the Interview

The interviews themselves were almost entirely unstructured in nature – while I had a broader set of potential questions to ask depending on their field of study and how our discussion progressed, there was no automatic pre-set question list. I chose unstructured interviews for several reasons. First, the problem with a structured interview in these cases is that I was not yet confident that I *knew* all the best questions

to ask. Second, given the wide variety of fields and approaches, it seemed unlikely that the insights provided would be of the same focus or easily comparable to one another. Rather, the intent was to obtain separate and distinct insights from different perspectives. Therefore, an unstructured interview design should be best.

5.5.4 Expert Interview & Peer Review Procedures

With the preceding subsections of 5.5 in mind, Stage Four involved the following specific steps:

1. Research for potential participants. This included reviewing scholars in the literature review section, simple web searchers for previous interviews conducted, as well as asking known scholars for recommendations.
2. Send initial communications to potential participants. Potential participants were contacted to ask if they would be interested in participating in this study, or if not, if they had any scholars in a similar field of study who might be interested.
3. If initial interest is identified, send more complete information as to what would be required of them, as well as obtaining their signatures for consent per the Institutional Review Board. *(Five participants were selected who met the selection criteria. The names of the five individual scholars can be found in Appendix C at the bottom.)*
4. Provide selected participants with a brief executive summary of my dissertation and my full draft analytical framework for their review.

5. Schedule a roughly hour-long interview with participants no less than four (4) weeks after receipt of my executive summary and draft analytical framework.
6. Conduct the interview and record the audio conversation for later perusal.
7. Update my analytical framework based on the discussions and critiques of my interviewees and write a new section identifying the key critiques from participants, my thoughts on their critiques, and whether or how those critiques were addressed or incorporated into the analytical framework
8. Send participants an updated draft of the analytical framework, along with the new section identifying the critiques provided from them and other participants. Optionally, participants could send me via email their final thoughts on the critiques and my responses to them.
9. Incorporate any final changes based on these replies and finish the second draft analytical framework

At the end of incorporating their changes, I created the *second draft analytical framework*, which can be found in *Appendix A-2*.

5.6 Stage Five: ATI Study Comparison and Finalize Analytical Framework

This stage of my methodology focused on the Alan Turing Institute's *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector* as a target for comparative analysis (and hereafter referred to as "the ATI Study" for ease of reference) (Leslie

2019). Of note, this stage in the methodology went beyond comparison for the sake of finding similarities and differences – it worked to actively build from and incorporate the best elements from the ATI study to make this analytical framework stronger.

The comparative analysis consisted of three parts. First, I identified the key differences in scope between the studies. Second, I went through each research thread (including splitting the different activities within democratic legitimacy into their own sections) and identified where the ATI study had relevant and/or competing key concepts. Finally, for each identified key concept, I summarized whether it merited inclusion as a new principle, enhanced an existing principle, or did not merit inclusion and why. It is also important to note that the same five selection criteria applied in 5.3.3 *Principle Assessment Criteria* during Stage Two are applied here for determining what qualified as a “Key Concept”.

5.6.1 Differentiating Comparative Analysis from Literature Review

While the ATI study could admittedly have fit within the literature review above instead of as its own Stage, I separated it for several reasons. First and foremost, it is the only study to my knowledge that qualifies as having attempted to create nearly the same kind of framework as this study does. Because of the sheer level of similarity, I argue that a significantly closer inspection is warranted from it rather than it simply being included as background information in the literature review.

Second, it doesn't easily fit into the structure of a literature review that is based on research threads because it at least touches on every research thread I cover. Unlike literature on algorithmic governance, the ATI study goes much deeper into how public agencies should use these ML systems on a practical level. Thus, it would require a fundamental change to the organizational structure of the literature review.

Finally, putting it at the end of the iterative stages of improvement for my framework allowed for a more effective and powerful one-to-one comparison between the ATI study and this study. Rather than using the ATI study to develop the fundamentals of how this framework should look from the beginning, the very fact that the ATI study and this study were developed entirely independent from one another (until this Stage) allowed for unique questions to be asked – for example, how did two analytical frameworks with similar intents but fundamentally dissimilar methodologies compare against one another in outcome? What do the differences between them mean? What can be learned from the ATI study?

6 Research Findings

With the research methodology defined, the research itself was conducted and documented. Below are the findings of my research.

6.1 Stage One Findings

The first stage involved bilateral comparisons of the different research threads. With six optimization problems each compared against five other optimization problems (and subtracting those not compared against explainability), this provided me with eleven pairs. For each pair, I identified which of the three bilateral relationships existed as defined in the Research Methodology section (namely, a Complementary Relationship, a Negative Relationship, or a Mixed Relationship). A visualization of these relationships is provided in *Section 6.1.4* below.

6.1.1 Democratic Legitimacy's Bilateral Relationships

First, I analyzed democratic legitimacy's relationship to each of the other threads. Democratic legitimacy is provided its own section for several reasons. First, it is arguably the most important research thread to consider for public agencies, as well as the thread through which each of the other threads intersect. Second, rather than base my arguments on the quantitative studies conducted by previous scholars, the bilateral

relationships of democratic legitimacy are argued based on the key legitimacy-inducing activities of democratic legitimacy defined in the literature review. Indeed, because the increasing or decreasing of democratic legitimacy cannot be defined on a quantitative basis, it does not easily lend itself to experiments or straightforward empirical evidence like the remaining bilateral pairs.

Below are my conclusions:

Accuracy <-> Democratic Legitimacy

This relationship is perhaps the simplest and most obvious. It should be self-evident that an ANN used in a public agency which is more accurate is inherently going to increase democratic legitimacy (while holding all other factors constant). At a minimum, accountability is enhanced when accuracy increases, as well as protecting one's due process rights.

Conclusion: Complementary

Fairness <-> Democratic Legitimacy

Fairness (except where otherwise noted referring to *algorithmic fairness*) is intrinsically critical to democratic legitimacy, particularly in the United States – however you define it, questions of fairness and bias permeate almost every major public policy process. Presuming the actual choice in how a public agency defined algorithmic fairness followed high standards of democratic legitimacy (itself a difficult and complex issue), the act of quantitatively increasing an ANNs algorithmic fairness would certainly

enhance substantive due process protections and accountability, as well as human autonomy and even non-algorithmic fairness.

Conclusion: Complementary

Explainability <-> Democratic Legitimacy

Making an ANN or other machine learning system more explainable should increase its transparency, interpretability, and accountability at the least. Transparency is enhanced when what is being made transparent isn't simply that "there is a black box" but rather "this is why we believe the not-as-black box made the decision that it did, and this is how we back up our reasoning". Likewise, it's difficult to hold a public agency accountable if there is no understanding of why a particular decision was made by an ANN. Due process rights are also protected when explainability is enhanced – it's difficult to know when an individual's rights are being violated if there is no explainability in an ANN's decisions, regardless of accuracy.

Conclusion: Complementary

Robustness <-> Democratic Legitimacy

Making an ANN more robust against malicious manipulation certainly enhances substantive due process rights – it is difficult to imagine one's constitutional rights being protected if the ANN ends up being manipulated into making incorrect decisions.

However, democratic legitimacy and robustness have a more complex relationship when it comes to transparency. The question relies on what, precisely, makes up robustness: does robustness extend only to the model itself, or is robustness *also* a function of the threat environment the model is in? In the former definition, transparency (or indeed any element of democratic legitimacy) have no impact on robustness since none of them directly modify the model itself. Regardless of how much transparency is provided, the model's architecture and weights are not modified in any way.

With the latter definition, however, robustness can be strongly and negatively impacted by increasing transparency. For example, if a public agency reveals all training and testing data and the model structure of the ANN they are using, this can make it significantly easier for malicious users to manipulate the outputs of that system. Indeed, the differences between “black box” and “white box” systems in robustness literature should make this point particularly clear – defeating the robustness of a black box system is significantly more difficult than a more transparent system (Alshemali and Kalita 2019). This does not mean that there shouldn't be transparency, but there needs to be a balance between the two.

For this study, I accept the latter definition of robustness which incorporates the threat environment as well as the model itself.

Conclusion: Mixed

Privacy <-> Democratic Legitimacy

Privacy is substantively like robustness in terms of its relationship to transparency. At first glance, increasing privacy (such as through differential privacy) is inherently complementary to democratic legitimacy – ensuring that the specific individuals used for training data cannot be recreated afterwards, for example, should only improve protections of due process rights and increase accountability and even human autonomy. However, enhancing transparency (and thus democratic legitimacy) can end up harming privacy through informing a malicious user about how a given ANN model was trained. With that information in hand, reverse engineering what elements were used to train the ANN (and thus the potential for the de-anonymization of data) should be easier. Indeed, (Young, et al. 2018) even provide a legal-technical framework for balancing the need for privacy with the need for democratic legitimacy (in this case accountability and transparency).

Conclusion: Mixed

6.1.2 Exclusion of Explainability

The only bilateral relationship which won't be further explored is that of explainability (except for qualitatively assessing it against democratic legitimacy above). This is because while explainability is highly important, it becomes extremely hard to

assess it as a positive or negative relationship with any of the remaining research threads. This difficulty stems from several areas.

First, the techniques discussed in the explainability section either (a) don't involve any direct change to the model itself, or (b) encompass techniques which improve explanatory power by happenstance (such as attention models) rather than by intent. Second, the sheer and ever-expanding quantity and variety of explanatory techniques (and the lack of consensus regarding which explanatory techniques are "best", which can easily vary with circumstance) makes it extremely difficult to assess in this way.

Finally, except for democratic legitimacy, explanatory power is the most qualitative research thread to assess. There is no definitive quantitative method of determining if a model is sufficiently explainable, as that will change with each situation. It can also be difficult at times to tell if something has become "more" or "less" explainable when comparing different methods. This contrasts with accuracy, robustness, fairness, and privacy, where the metrics for whether they have become more or less optimized is almost entirely quantitative in nature, even if there is debate over the proper quantitative metric.

6.1.3 Other Bilateral Relationships

Therefore, aside from democratic legitimacy's bilateral relationships and the exclusion of explainability, there are six other relationships that need to be explored:

1. Fairness vs. Accuracy
2. Robustness vs. Accuracy
3. Privacy vs. Accuracy
4. Robustness vs. Fairness
5. Privacy vs. Fairness
6. Robustness vs. Privacy

Unlike with democratic legitimacy, where evaluation is almost entirely qualitative in nature and which made practical testing next to impossible, each of these relationships below are based on at least one piece of empirical scholarship. Scholarship was selected if it met the following conditions:

- The scholarship conducted actual quantitative experimentation which concluded one way or the other as to the nature of the bilateral relationship, even if that experimentation was not meant to be generalizable to all use cases
- If the experimental testing was not with an ANN, then the scholarship considers ANNs and the potential differences between their own study and whether their conclusions should be relevant to ANNs
- The scholarship was available for review no later than December 1st, 2019

Nevertheless, some relationships only have minimal empirical scholarship available, some had literature only on the pre-print server arXiv and was not yet peer-reviewed, and as mentioned previously, some literature did not set out to be generalizable for the entirety of the relationship. Because of this, the conclusions reached in this section should be taken as highly preliminary and may be subject to change as future scholarship becomes available. Indeed, only the strongest conclusions from this section are included in the final analytical framework.

Fairness <-> Accuracy

No matter what fairness definition is used, current empirical literature shows that fairness constraints hamper accuracy, or at absolute best do nothing to improve it (Wadsworth, Vera and Piech 2018) (Raff and Sylvester 2018) (Jagielski, et al. 2019) (S. Friedler, et al. 2018) (Yurochkin and Bower 2019).

According to (Raff and Sylvester 2018), “[i]t would be unusual to expect adding the fairness constraint to any classifier would significantly *increase* accuracy.” Additionally, (Jagielski, et al. 2019, 17) show that as fairness increases (defined by the authors as *equality of odds*), accuracy is likely to decrease. This is one of the strongest and most consistent relationships identified in empirical literature.

Consensus: Tension

Robustness <-> Accuracy

For the purposes of this study, robustness is defined as *resiliency to adversarial examples*; when robustness is defined in terms of *label noise*, some scholarship has actually shown a positive relationship between accuracy and robustness (Vahdat 2017) (Hendrycks and Dietterich 2019). However, the relationship between adversarial examples and accuracy is complex – currently available literature provides some empirical evidence indicating that while optimizing for robustness with a small training set may actually *increase* accuracy, optimizing for robustness with a larger dataset will likely *decrease* accuracy (although not all scholarly literature tested models based on

smaller training datasets) (Tsipras, et al. 2018) (Su, et al. 2019) (Lei, Wang and Su 2019) (Zhang, et al. 2019). Current scholarship theorizes that this is because with lower quantities of data, the decision boundaries are drawn too sharply and optimizing for robustness can cause these decision boundaries to be more blurred. With higher quantities of data, the decision boundaries are already sufficiently blurred to maximize accuracy and any further blurring causes false positives or false negatives, which thus decreases accuracy. Therefore, I consider it a mixed relationship – there are so many factors which influence this relationship (not the least of which is one’s definition of robustness itself) that it can easily vary between ANNs.

Conclusion: Mixed

Privacy <-> Accuracy

Current literature is nigh-unanimous in concluding that optimizing for differential privacy will decrease accuracy, at least to some degree (Shokri and Shmatikov 2015) (Yu, et al. 2019) (Bagdasaryan and Shmatikov 2019) (Phan, Thai, et al. 2019) (Jayaraman and Evans 2019). Indeed, some have taken the argument further and linked it to fairness: for example, (Bagdasaryan and Shmatikov 2019) assert that for smaller groups in a dataset (i.e. a racial or ethnic minority) “accuracy of DP [differential privacy] models drops much more for the underrepresented classes and subgroups.”

Conclusion: Tension

Robustness <-> Fairness

While there was no literature I was able to find comparing robustness and (algorithmic) fairness in broadly applicable terms to all ANNs, there is some literature focusing explicitly on the intersection of fairness and robustness in text classification (Garg, et al. 2019) (Yurochkin and Bower 2019). Although the authors have different methodologies, the basic premise of their papers is the same: they attempt to achieve fairness through providing robustness.

While (Garg, et al. 2019) explicitly focus on individual fairness, (Yurochkin and Bower 2019) work to link their scholarship closer to group fairness. Both cases are also focused on specific use cases for text classifiers, which may not be broadly applicable to other uses. Both of the authors sought as their ideal that there should be little to no change in a machine learning system's output if the only change is to specific "protected words" (i.e. words dealing with a protected subgroup) in the text.

For example, a machine learning system assessing an applicant's resume for a job would be fair under both of their definitions if the score it provided didn't change regardless of if the applicant's name is likely to be Caucasian or African-American, or for differing genders. Both papers show that by making their system robust to such changes, they achieve fairness as they define it. Although their definitions of fairness are somewhat narrow and their use case is hard to generalize from, they do appear to show a positive relationship between robustness and fairness.

Conclusion: Complementary, but with caveats

Privacy <-> Fairness

Although this literature is still young (and at the present only focuses on differential privacy), the current consensus in existing scholarship is that privacy and fairness are in tension with one another (Bagdasaryan and Shmatikov 2019) (Jagielski, et al. 2019) (Cummings, et al. 2019). First, (Bagdasaryan and Shmatikov 2019) assert that “if the original model is unfair, the unfairness becomes worse once DP [differential privacy] is applied. We demonstrate this effect for a variety of tasks and models, including sentiment analysis of text and image classification.” What is more, the authors show that the damage to fairness is even worse for underrepresented minority groups. Likewise, (Jagielski, et al. 2019) appear to agree with this assessment: while they showed that they could achieve realistic differential privacy with a relatively small tradeoff in accuracy and fairness, the fact that tradeoffs existed is still true (Jagielski, et al. 2019, 17). Finally, (Cummings, et al. 2019) assert that while it is mathematically impossible to achieve differential privacy with *exact* fairness and “non-trivial” accuracy, they also try to prove mathematically that what they define as “approximate fairness” can be achieved alongside differential privacy with low cost. Nevertheless, while such research in minimizing the tension is certainly useful, the fact that the two threads remain in tension by default remains.

Conclusion: Tension

Robustness <-> Privacy

Existing literature focused on this relationship is nigh-unanimous that there is a complementary (or at least neutral) relationship between optimizing for differential privacy and robustness (Lecuyer, Atlidakis, et al. 2018) (Phan, Vu, et al. 2019) (Phan, Thai, et al. 2019) (Lecuyer, Atlidakis, et al. 2019). This makes intuitive sense as well, since both methods semantically focus on the same problem: an adversary is attempting to take advantage of the machine learning system in some way, and the designer must thus try to mitigate this issue.

Conclusion: Complementary

6.1.4 Summarizing Bilateral Relationships

From these findings, I have created the following summary table for reference:

Table 8 - Bilateral Relationships of Research Threads

Research Thread	Accuracy	Privacy	(Algorithmic) Fairness	Robustness	Explainable AI
Accuracy					
(Algorithmic) Fairness					
Robustness			*		
Privacy					
Democratic Legit.					
*Caveats apply					

Key:

Red: Tension relationship

Green: Complementary relationship

Yellow: Sometimes complementary, sometimes tension

While such a simplified form as this table may lack nuance and is a valid target of critique because of this, I believe it nevertheless captures enough of these relationships to be a worthwhile visual addition.

6.1.5 Beyond Bilateral Relationships

While analyzing these bilateral relationships is the primary purpose of this stage of my methodology, it is also important to look where scholars are pushing research even further. One of the most powerful pieces of recent scholarship is (Sharma, Henderson and Ghosh 2019), where they produce a model-agnostic auditing system for ANNs and other ML models. Their auditing model uses the concept of counterfactuals to examine not only accuracy, but also “robustness, interpretability, transparency, and fairness.” While their current model doesn’t assess for privacy, it should nevertheless be seen as an important stepping stone to the kind of methodical, standardized testing framework necessary for implementing ANNs in a public policy setting.

6.1.6 Conclusions

In conclusion, while I believe the first stage has value in terms of assessing the current state of comparative research between these different threads, it is important not to draw too much from its conclusions. There is simply not enough research, and particularly not enough broad and conclusive research, to definitively prove that all the relationships identified above are constant throughout all ANNs and machine learning systems and in all situations.

Rather, they should be seen as a starting point for understanding the current state of research and show potential avenues for more broader studies as to these relationships. I do not doubt that the relationships identified above may change or even be proven wrong once there is deeper research.

6.2 Stage Two Findings

This section includes the principles extracted from one or more of the existing AI frameworks specified in *Section 5.3* above. Each extracted principle includes the framework (or frameworks) that the principle was gleaned from. Once extracted, some of the relevant principles were merged with similar principles or split into separate principles for the first draft analytical framework; the first draft can be found in *Appendix A-1*.

For the sake of clarity and focus, only those principles which were at least on the borderline of meeting all five criteria (the idea of *borderline* denoted below as a yellow caution sign) are included. Otherwise, there would be hundreds of additional principles from the ethical AI frameworks above to add simply to be immediately rejected. Common types of rejected principles include those related to AI legislation advocacy (failed Criteria 2), generic statements about not violating a given right (failed Criteria 3), and principles related to how to regulate private sector AI usage (failed Criteria 2 & Criteria 5).

As I have noted previously, the purpose of this study is not to assert that ML systems should or shouldn't be used for particular policy end goals or that a particular kind of legislation or regulation would be ideal for managing such systems. Rather, it is to find those normative procedural principles that are most important for successful development and implementation, regardless of what one's policy goals are.

To restate, here are the five criteria I use from *Section 5.3.3* above:

Criteria 1: The principle is not just a computer science principle

Criteria 2: The principle is relevant to US domestically-focused public agencies

Criteria 3: The principle is not overly generalized or self-evident

Criteria 4: The principle does not violate democratic legitimacy

Criteria 5: The principle does not rely on normative assertions of the results of specific policies

6.2.1 Summary of Extracted Principles

Some frameworks simply did not contain any principles besides generic ones. Indeed, most potential principles from the government grand strategy documents rarely moved beyond generic principles and thus failed to meet Criteria 3. Here are those principles that were accepted which were at least on the borderline for all five criteria:

Table 9 - Extracted AI Principles

	Citations	Criteria 1	Criteria 2	Criteria 3	Criteria 4	Criteria 5*
Principle						
A human must always be in	(UNI Global Union 2017) (The Public					

control of an AI's decisions	Voice 2018) (Google 2019) (Coglianese and Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era 2017) (Abrassart, et al. 2018) (Future of Life Institute 2017) (High-Level Expert Group on AI 2019)					
Ban attribution of responsibility to robots	(UNI Global Union 2017) (Abrassart, et al. 2018)	✓	✓	✓	✓	✓
Conduct external impact assessments	(Amnesty International 2018) (Reisman, et al. 2018)	✓	✓	✓	✓	✓
Disclose known vulnerabilities	(Amnesty International 2018)	✓	✓	✓	✓	✓
Publicly disclose where systems are being used	(Amnesty International 2018) (The Public Voice 2018)	✓	✓	✓	✓	✓
Avoid 'black box systems'	(Amnesty International 2018)	✓	✓	⚠	✓	✓
Utilizing a contractor negates none of the public	(Amnesty International 2018)	✓	✓	✓	✓	✓

agency's responsibilities						
Use a human-centered design approach	(Pásztor 2018) (Lovejoy 2018)	✓	✓	✓	✓	✓
Empower users to test themselves	(Pásztor 2018)	✓	✓	✓	✓	✓
Differentiate AI content visually	(Pásztor 2018)	✓	✓	✓	✓	✓
Make explainability visual where possible	(Pásztor 2018)	✓	✓	✓	✓	✓
Ensure users understand their role in calibrating a given system	(Lovejoy 2018)	✓	✓	✓	✓	✓
If a human cannot perform a task, neither can an AI	(Lovejoy 2018)	✓	✓	⚠	✓	✓
Solicit public comment prior to implementation	(Reisman, et al. 2018)	✓	✓	✓	✓	✓
Correlation does not equal causation, particularly with ANNs	(Anastasopoulos and Whitford 2019)	✓	✓	✓	✓	✓
Do not assign responsibility to a "lower-level analyst"	(Coglianese and Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era 2017)	✓	✓	✓	✓	✓

Usage of ML systems may force previously qualitative values to be quantified	(Coglianese and Lehr, Regulating by Robot: Administrative Decision Making in the Machine-Learning Era 2017)					
Critically important final decisions regarding a person's life, quality of life, or reputation must be made (time and circumstance permitting) by a human being	(Abrassart, et al. 2018)					
It is legitimate to restrict access to an AI's algorithm and training data when there is a high chance of endangering public health or safety.	(Abrassart, et al. 2018) (Trump 2019)					
AI-generated user behavior profiles should be treated with great caution	(Abrassart, et al. 2018)					
People should have the right to access, manage and	(Future of Life Institute 2017)					

control the data they generate						
Determine if special redress procedures are necessary for the machine learning system's determinations	(Shrum, et al. 2019) (Reisman, et al. 2018)	✓	✓	✓	✓	✓
Determine if the machine learning system is engaging in "nudging", and if so, how acceptable such nudging is.	(Shrum, et al. 2019)	✓	✓	✓	✓	✓
Vulnerable groups, such as the differently abled, children, the elderly, the poor, and disadvantaged minority groups, are at particularly acute risk from ML systems	(High-Level Expert Group on AI 2019)	✓	✓	✓	✓	✓
Consider both technical and non-technical methods to ensure that problems that may arise from ML systems are solved	(High-Level Expert Group on AI 2019)	✓	✓	✓	✓	✓

6.2.2 Discussion on “Arguable” Principles

In the section above, there were three principles identified which were on the borderline of meeting certain criteria. They are discussed in greater depth in this section as to their inclusion or exclusion:

Avoid using ‘black box’ systems

While in an ideal world no ANN or machine learning system would be a black box, at present it is nigh-impossible to find an ANN (or an advanced ML system) that isn’t at least *partially* a black box. While attempting to mitigate this black box problem through explainability is certainly important (indeed, an entire section of the literature review is based on how to do this), a blanket assertion that such systems should never be used is overly prescriptive at the least.

Depending on how one defines a black box, including such a principle could preclude usage of any and all ML systems, regardless of their potential benefit. While the fact that these systems are often at least partially black boxes is a reasonable concern, it should not be an overriding one in all cases regardless of the circumstances. A more moderate version of this principle will be included in the framework instead which outlines the potential harms while also leaving such decisions to the public agency manager to weigh costs and benefits.

If a human cannot perform a task, neither can an AI

The problem with this principle is that it requires a significant amount of context, and its potential to be used out of context (and thus drawing potentially harmful

conclusions from it) is high. This principle is specifically meant in the context of applying labels to training data – if human beings cannot successfully label training data, then a supervised machine learning system can't be expected to learn the task. This is undoubtedly true, and a valuable point to consider, but if one applies this principle too broadly it can lead to highly fallacious conclusions. For example, no human being is capable of quickly and correctly identifying one face from a database of millions, yet ANNs can do this with ever-increasing accuracy. In short, this principle's potential for misunderstanding and misapplication require it to be substantially reworded for inclusion

People should have the right to access, manage, and control the data they generate

This is a broad principle that arguably crosses the line into policy outcome versus procedure. However, it encompasses a far more fundamental discussion of data, privacy, and government: how much control should an individual have over data the government legally collects from them? Should every individual have the right to opt-out of having their data in any public agency's machine learning system? Or to go even further, should public agencies be by default forbidden to use people's data in a machine learning system without the express permission of Congressional legislation? Or can the risks to privacy and human autonomy be sufficiently mitigated to allow it in some circumstances?

While this principle encompasses procedures as well as outcome, I believe it pushes too far into subjective political outcomes for it to be included in this framework.

This is not to indicate that the question is unimportant, but rather that it is not meant to be answered by this study.

6.2.3 Conclusion

With a set of principles extracted from the ethical AI frameworks (along with the findings from Stage One), a first draft analytical framework was constructed; it can be viewed in *Appendix A-1*.

6.3 Stage Three Findings

From the first two stages, there are three basic categories of principles which were created for organizational purposes. These categories may need to be further expanded and changed as this methodology changes and evolves, but they provide a sufficient place to begin with.

6.3.1 Initial Categories of Principles

There are three categories of principles in the first draft analytic framework. First, there are the *Optimization Principles*. These principles will be predominantly (though not entirely) gathered from the Stage One findings covering how different research threads relate to each other and the problems that can be faced when trying to optimize multiple principles simultaneously. Second, there are *Human Interaction* principles. These principles are largely drawn from Stage Two and cover those areas where the human-machine interaction is the primary element of study. Finally, there are

General Principles. This is a catch-all category for principles which don't easily fit into either of the two preceding categories and come from either Stage One or Stage Two.

6.4 Stage Four Findings

Stage Four consisted of five interviews combined with five concurrent peer reviews from individuals in a variety of related fields. While my ideal would have been to have 8-11 such interviews/peer reviews completed (as is common in qualitative interviewing in the social sciences), this was unfortunately impossible to accomplish. There are thousands of scholars in the United States focused on machine learning and ANNs, but unfortunately the number that (a) had an interdisciplinary focus between computer science and the social sciences, and (b) agreed to participate, was much smaller. Over 80 scholars were contacted, but only 5 participated to the end of the process.

Before I provide an updated framework, it is important to specify the core critiques presented over the course of these expert interviews and peer reviews. To ensure that participants were able to speak freely about their views, specific quotations or attributions of specific critiques to one or more individuals are not provided. While most critiques presented were accepted to varying degrees, I provide accompanying explanations as to my reasoning why or why not a given critique was accepted and how it was integrated into the framework.

From the interviews and peer review, several broad ideas emerged. Most of those ideas impacted the analytical framework itself, although some also caused additions to the literature review or introduction sections as well. Where areas outside the analytical framework itself were impacted, they are also specified.

6.4.1 Key Critiques

The second draft analytical framework (that is, after completing the interviews and attempting to mitigate the critiques it provided but before conducting the comparative analysis in Stage Five) is *in Appendix A-2* at the end of the study. It is not provided in this section to avoid confusion between the drafts and the final version of the framework. However, the key critiques of the interviewees are listed immediately below, along with a summary of the changes implemented from these critiques into the second draft.

Critique 1: Integrate principles aimed at the public agency's relationship with the software vendor

One major element argued to be missing from the original draft framework was what principles should be used when dealing with software vendors. While some rare public agencies might have the ability in-house to create such internal machine learning tools, current research agrees that the large majority rely on external vendors for their software solutions, particularly in the case of machine learning systems (Shrum, et al. 2019) (Reisman, et al. 2018) (Brauneis and Goodman 2018) (Ram 2017) (Wexler 2017).

Table 10 - Summary of Actions Taken 1

Summary of Actions Taken	
Implementation:	Full
Improvement(s):	- Added new Vendor section to analytical framework - Re-reviewed findings from Stage One and Stage Two for relevant vendor-oriented principles that may have been overlooked.

Critique 2: Focus on Machine Learning more broadly instead of just Artificial Neural Networks

This was an idea mentioned by several interviewees: since most of the principles are applicable (to varying degrees) to machine learning systems in general, the focus on artificial neural networks should be reduced or even removed. This idea was partially implemented. While the dissertation is still focused on ANNs more than other kinds of machine learning, significant additional attention was given to machine learning in general. Additionally, added explanation is provided as to what machine learning is defined as and what kinds of techniques fall under machine learning as defined in this study.

Additionally, ideas or principles that are relevant to more than just ANNs were changed to machine learning in the text throughout the dissertation, and additional literature was added to the literature review as well. For the analytical framework itself, each principle was specifically identified as applying to either ANNs specifically or to machine learning systems more generally (either None, Partial, or Full). Finally, previous

sections delineating what methods constituted machine learning and what methods did not were updated and refined.

Table 11 - Summary of Actions Taken 2

Summary of Actions Taken	
Implementation:	Partial
Improvement(s):	- Literature review expanded - Definition of machine learning and its relationship to ANNs refined - Draft framework principles refined to better exemplify whether they were meant as applicable to ANNs or ML at large.

Critique 3: Democratic Legitimacy is too vague and undefined to argue for normatively
 It is not immediately clear at face value what achieving democratic legitimacy entails to be proven or disproven. Because of this, it was originally suggested that a different principle such as social welfare be chosen instead, or alternatively that democratic legitimacy be better defined so that the reader understands what is being considered. I chose the latter option, significantly refining what actions were required to achieve democratic legitimacy.

To make these determinations, I first added more *algorithmic governance* literature within my literature review section on democratic legitimacy. Then, I worked to integrate the ideas from traditional democratic legitimacy literature and algorithmic governance literature to determine what the key activities for achieving democratic

legitimacy are in the context of machine learning systems. Finally, I added a new element to each of my framework's principles specifying if it was relevant to achieving democratic legitimacy, and if so, which specific activities within democratic legitimacy.

Table 12 - Summary of Actions Taken 3

Summary of Actions Taken	
Implementation:	Partial
Improvement(s):	- Algorithmic governance literature added to literature review section under democratic legitimacy - Definition of the activities that make up democratic legitimacy refined to integrate concepts from algorithmic governance - Added relevance to democratic legitimacy activities (if applicable) to each principle

Critique 4: Remove overly vague principles

This is a catch-all for various several specific criticisms of my original draft framework's principles. Four principles were entirely removed due to vagueness. I agreed with these critiques, and removed the offending principles:

- General Principles #1, #6:
- Optimization Principles #2, #5

Table 13 - Summary of Actions Taken 4

Summary of Actions Taken	
Implementation:	Full
Improvement(s):	- Principles specified above were removed or combined with other principles

Critique 5: Insufficient focus on the outcome versus the process

This is a valid critique and one I accept. It comes down to what the focus of this study and this analytical framework is intended to be. This framework is not meant as a guide to judge a particular policy outcome as normatively “good” or “bad”. Rather, as I have noted several times, the framework is meant to be *procedurally* normative in focus: what are the best principles to consider for designing and maintaining a machine learning system, regardless of one’s desired policy goals. This also connects to the focus on helping users to ask the right questions rather than answering those questions. However, this point was not clear enough in the dissertation previously and has now been made clearer. The idea was minimally implemented to increase clarity as to this study’s focus.

Table 14 - Summary of Actions Taken 5

Summary of Actions Taken	
Implementation:	Minimal
Improvement(s):	- Added additional clarity to Introduction, Research Methodology sections specifying that the normative

	focus of this dissertation is procedural and not on specific policy outcomes.
--	---

Critique 6: Tie principles together with additional structure

One critique was that the principles did not “tell a story”, but rather were just disjointed bits of knowledge. Even if they were valuable data points individually, they were not sufficiently connected to be considered a framework or to be effectively understood in relationship to the other principles.

This idea was fully accepted. While I had initially concluded that categorizing the principles was enough to connect them to one another, there remained a lack of connection between principles of different categories and indeed at times between principles of the same category. What is more, some principles arguably fit under two categories simultaneously.

To combat this, I added a summary table to the beginning of each principle. Each principle’s summary table would show which specific research thread(s) it was related to, as well as whether or not the principle had a secondary category it could be simultaneously placed in and whether the principle had one or more other principles that were most closely related to it.

Table 15 - Summary of Actions Taken 6

Summary of Actions Taken	
Implementation:	Full
Improvement(s):	- Summary table added at the beginning of each principle showing which specific research thread(s) it relates to, as well as whether the principle has a secondary category.

Critique 7: Split General Principles category

One critique of the original General Principles section was that it was simply too inclusive – it included both highly technical principles as well as principles that were broader or more overarching. In order to alleviate this issue, I split the original General Principles section into General Sociotechnical Principles and Public Agency Manager Principles. While the former category is still something of a catch-all, the principles within them both are at least somewhat more in line with one another than previously. This idea was fully implemented.

Table 16 - Summary of Actions Taken 7

Summary of Actions Taken	
Implementation:	Full
Improvement(s):	- Split General Principles into General Sociotechnical Principles and Public Agency Manager Principles

Critique 8: Better define and explain the differences between Transparency, Interpretability, Accountability, and Explainability

This critique was based on my original focus on explainability as a catch-all for all four of these concepts when that is not necessarily the case. This is a valid point: while the four concepts are strongly related to each other, they are not necessarily equivalent or even entirely overlapping. For example, a machine learning system can be transparent but lack explainability: even if all input data is provided as well as the technical details such as weights and neurons, this does not automatically imply that the system is explainable. Additionally, the idea of transparency also applies to the structure of the public agency itself and their rules on releasing that data to the public. Even if the machine learning system is explainable internally, that does not mean that this explainability will be released to the public.

Likewise, accountability and explainability can exist independently of one another: even if a machine learning system is sufficiently explainable, the concept of accountability extends much further than the system itself. Rather, accountability encompasses the governing structure of the public agency that surrounds the machine learning system; it is not something that can be optimized for by modifying the structure of the machine learning system itself.

Table 17 - Summary of Actions Taken 8

Summary of Actions Taken	
Implementation:	Full
Improvement(s):	- Added Accountability, Interpretability, and Transparency as unique concepts within Democratic Legitimacy

6.4.2 Summary of Key Improvements from First to Second Draft of Analytical Framework

In summation, the following key improvements were made from the first draft of the analytical framework to the second draft:

- General Principles category split into Public Agency Manager Principles and General Sociotechnical Principles
- New Vendor Principles category added
- Refined definition of the relationship between ML systems and ANNs
- Added key references to algorithmic governance literature
- Added summary table for each principle to assist in linking the principles together
- Greatly enhanced the definition of democratic legitimacy by adding greater specificity to legitimacy-seeking activities
- Added net total of eight principles from first draft to second draft, including the removal of several principles deemed overly vague

6.5 Stage Five Findings

With each of the preceding stages complete, I had the second draft analytical framework ready, which can be viewed in *Appendix A-2*. From this point, Stage Five involved a comparative analysis of my framework against the ATI study's framework. Rather than simply include the ATI Study in the literature review (or as a subset of ethical AI literature), I concluded that it was similar enough in concept to warrant its own Stage of analysis for iterative improvement of my own draft analytical framework.

6.5.1 Identifying Key Structural Differences

This study focused on six research threads: accuracy, explainability, fairness, robustness, privacy, and democratic legitimacy. In contrast, ATI structured their principles differently. While both studies attempt to provide guidance to public agency managers in dealing with AI systems, there are nevertheless several key differences in their scope that bear closer examination.

National Focus

First, both this study and ATI's have a different national focus. Although the ATI study itself does not explicitly endorse usage for public agencies in a particular country, the Alan Turing Institute itself is the United Kingdom's national institute for data science and artificial intelligence. As such, it is reasonable to construe their framework as being oriented towards public agencies in the UK. By contrast, this study focuses on US public agencies.

AI Ethics

Second, both studies focus on AI ethics to differing degrees. This study focuses more exclusively on procedurally normative best practices for public agency managers. In contrast, ATI's study attempts to discuss what ethics and morality themselves entail in both procedure and outcome.

AI vs. ML vs. ANNs

Third, both studies have a somewhat different focus in terms of artificial intelligence vs. machine learning vs. artificial neural networks. Although ATI's study uses the term "AI", it is clear that they are referring to machine learning rather than symbolic AI systems. Indeed, the implication is present from their very first page of content that they are specifically interested in "increasingly sophisticated machine learning algorithms" (Leslie 2019, 3). While this is not *precisely* the same as the focus of this study, it is nevertheless close enough to allow for a comparative analysis without significant hurdles.

Near Term vs. Long Term

While this study is predominantly focused on proper design and initial implementation of for ANNs (and ML systems more broadly), ATI's study theorizes over the administrative mechanics of public agency implementation over the mid to long term.

6.5.2 Limitations of the ATI Study

While building on ATI's work through a comparative analysis is expected to improve this study's analytical framework, ATI's study is nevertheless lacking in two

specific areas of content: methodically dealing with research thread conflicts and dealing with software vendors.

First, while it does mention the concept of trade-offs several times, it does not develop the idea too deeply. The ATI study discusses trade-offs in ethical values generally (Leslie 2019, 11), trade-offs in algorithmic fairness (Leslie 2019, 18), and trade-offs between accuracy and interpretability (Leslie 2019, 44). However, it lacks a more methodical analysis of recent literature as to what additional tensions might exist (for example, between privacy/robustness and accuracy). Additionally, it provides little information as to which research threads do the opposite – where improving optimization in one thread has a complementary (or mixed) relationship with another.

Second, it does not touch on external vendors. In the US especially, it is a rare event when a domestically focused public agency can entirely implement its own ML models without outside vendor assistance. While it is possible that UK government agencies have substantially more internal technical expertise and require less from vendors, I find this unlikely to be the case. Unfortunately, comparative information about the frequency of utilizing outside contracting firms for developing machine learning systems between the US and UK governments is not presently available.

6.5.3 Differing Conceptions of the Key Research Threads

Another difference was how both studies conceived of the different research threads. It is important to delineate this since there are situations where both this study

and the ATI study use the same terminology to refer to a potentially different concept. This is not to say that the ATI study's conceptions are wrong, but simply that they may not be referring to the same thing.

First, the ATI study more closely aligned algorithmic fairness and non-algorithmic fairness under a single fairness umbrella, rather than splitting them between "algorithmic fairness" and "non-algorithmic fairness". More specifically, the ATI study explains non-algorithmic fairness in terms of data fairness, design fairness, and implementation fairness (although some elements of these concepts arguably fall outside of even non-algorithmic fairness as this study defines them), whereas the ATI study uses outcome fairness when they are referring to algorithmic fairness. This will be discussed in greater detail below.

Second, the ATI study does not always make clear the differences between explainability and interpretability. While it goes into detail to define interpretability, it does not signify how (or if) the idea of explainability is different than interpretability. Rather, it appears to use them relatively interchangeably.

Third, it defines privacy as more of an organizational issue rather than as a technical issue. While it does not mention concepts such as differential privacy, homomorphic encryption, or federated learning, its conception of privacy is much more firmly related to individual rights and invasions of privacy by public agencies and

malicious actors. In this study, those ideas are captured within democratic legitimacy and the concept of privacy is left to ensuring privacy at the algorithmic level.

With these in mind, a review of each of this study's research threads and how the ATI study's key concepts relate to them is below. The conclusion section(s) for each research thread below are italicized. In areas where additional summarizing information is needed for a key concept, it will be included above the conclusion section. When multiple Key Concepts are similar enough that they can be analyzed with a single conclusion section, they are placed immediately after one another.

6.5.4 Accuracy

Key Concept 1: Model accuracy is often based on complex social/historical patterns which may contain encoded bias from cultural norms

Key Concept 2: Accuracy is not necessarily an absolute metric based on the 'ground truth' – it can be subject to biased human decisions and judgments when the data is subjective. (Leslie 2019, 14-15)

Conclusion: Both key concepts in this section are already somewhat covered under General Sociotechnical Principles, Principle 3. However, they both add additional context and bring greater clarity to the meaning of the principle. As such, this principle will be updated to include this information.

6.5.5 Fairness (non-Algorithmic & Algorithmic)

Like accuracy, fairness was a key subject of focus in both the ATI study and this one. Indeed, both studies attempted to spell out various definitions of fairness (Leslie 2019, 13-22).

Key Concept 1: Data Fairness

The ATI study defines *data fairness* as making sure a machine learning system is “trained and tested on properly representative, relevant, accurate, and generalisable datasets” (Leslie 2019, 14). This type of fairness is most closely related to this study’s definition of “non-algorithmic fairness” within democratic legitimacy.

Conclusion: Updates to General Sociotechnical Principles, Principle 3 from the Accuracy section above should have covered ensuring this concept is integrated.

Key Concept 2: Design Fairness

Design fairness is considered through making sure that the “model architectures...do not include target variables, features, processes, or analytical structures (correlations, interactions, and inferences) which are unreasonable, morally objectionable, or unjustifiable” (Leslie 2019, 14).

Conclusion: This concept is alluded to in this study generally and General Sociotechnical Principles, Principles 6 & 7 touch upon it, but there is still more that can be said. A new principle will be added to General Sociotechnical Principles discussing it in greater depth – while the ML system itself decides which variables are “important”, it is still up to human beings to decide which variables are selected as possibilities in the first place.

Key Concept 3: Outcome Fairness

Outcome fairness is what is meant by the broader Fairness section of the literature review. Below is a basic mapping of the fairness concepts that the ATI study considers, and a comparison to how this study identifies and considers them:

Table 18 - Outcome Fairness Study Comparison

ATI Study	This Study	Notes
Demographic/ Statistical Parity	Parity (Wadsworth, Vera and Piech 2018)	<i>Parity</i> is essentially the same as ATI's conception for demographic / statistical parity
True Positive Rate Parity	Equality of Odds (Wadsworth, Vera and Piech 2018)	<i>Equality of Odds</i> achieves both True Positive and False Positive Rate Parity
False Positive Rate Parity		
Positive Predictive Value Parity (PPVP)	Equal <i>precision</i> across groups	This standard of fairness is not explicitly referenced in this study, but it is implicitly identified through the <i>precision</i> metric – PPVP refers to the case where precision is equivalent across protected groups.
Individual Fairness	Basic individual fairness	This is the basic definition for individual fairness covered in the Fairness section of the literature review.
Counterfactual Fairness	N/A	“Counterfactual fairness” is as much about explainability as it is about fairness. While is not covered in the Fairness part of the literature review, it is covered briefly in Section 6.1.4 above.

Conclusion: These conceptions of fairness are for the most part already considered in this study. No new information is updated.

Key Concept 4: Implementation fairness

According to the ATI study, implementation fairness is when machine learning systems are “...deployed by users sufficiently trained to implement them responsibly and without bias”. This includes both *Decision-Automation Bias* and *Automation-Distrust Bias*. Decision-automation bias is when the user is “...hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the perceived objectivity, neutrality, certainty, or superiority of the AI system” (Leslie 2019, 21-22).

At the opposite end of the spectrum, automation-distrust bias is when the user “disregard[s] its [the machine learning system’s] salient contributions to evidence-based reasoning either as a result of their distrust or skepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise” (Leslie 2019, 21-22).

Conclusion: While issues similar to these coming from the field of behavioral psychology were identified at the end of the literature review in Section 4.7.3, they were not originally included in the framework itself since behavioral psychology was not a chosen research thread. However, this section makes a compelling case for their inclusion. A new principle will be added to Human Interaction Principles section.

6.5.6 Explainability

While the ATI study is somewhat vague on where it draws the line between explainability and interpretability, it nevertheless has several key concepts based on the idea.

Key Concept 1: Lack of explainability may be acceptable in some cases, although conflicts with issues of fairness/discrimination make it more potentially problematic. (Leslie 2019, 4)

Conclusion: This concept is already covered within Public Agency Manager Principles, Principle 1.

Key Concept 2: When machine learning systems draw from human relationships and social patterns for data, designers must ensure there is enough interpretability such that the systems are non-discriminatory. When this is not possible, a more transparent and explainable model should be chosen. (Leslie 2019, 17)

Conclusion: This concept should already be covered in previously implemented additions to General Sociotechnical Principles, Principle 3.

Key Concept 3: There are four explanatory strategies: internal explanation, external explanation, supplemental explanatory infrastructure, and counterfactual explanations. (Leslie 2019, 50)

Let us first compare these conceptions to the conceptions described in the literature review's explanatory techniques taxonomy. While they do not perfectly match up, they cover the same general families of techniques:

Table 19 - Explanatory Techniques Study Comparison

ATI Study	This Study	Notes
Internal Explanation	Representation Techniques Disentanglement Techniques	N/A
External Explanation	Processing Techniques	N/A
Supplemental explanatory infrastructure	Attention Techniques Generated Explanations	N/A
Counterfactual explanations	Processing Techniques Democratic Legitimacy (Interpretability)	Counterfactual explanations are a type of external explanation that also encompasses increasing the interpretability of the model (Leslie 2019, 50).

Conclusion: No additional information needs to be added.

6.5.7 Robustness

While the ATI study writes a lot on robustness, most of it relates to defining adversarial attacks and data poisoning, as well as summarizing technical methods of defending against them, rather than prescriptive ideas to help optimize robustness.

Key Concept 1: The key risks that robustness is meant to counter are adversarial attacks, data poisoning, and misdirected reinforcement learning behavior (Leslie 2019, 32-34)

Conclusion: Both adversarial attacks and data poisoning are covered in the robustness section of the literature review; the third element to this concept is arguably related to data poisoning and is too niche to consider granting its own principle. To ensure that

framework does not ignore the point, a new question was added to Public Agency

Manager Principles, Principle 1.

6.5.8 Privacy

Key Concept 1: When dealing with stakeholders during project formulation, public agencies should determine how a given AI system might infringe on privacy rights both in terms of system design and system deployment.

(Leslie 2019, 28)

Conclusion: Stakeholder discussions are already included in Public Agency Manager

Principles, Principle 4, and Vendor Principles, Principle 4. Additional commentary related to privacy from this concept will be added to the latter.

6.5.9 Democratic Legitimacy (Transparency)

Key Concept 1: Be able to justify your process

Key Concept 2: Be able to explain results in a clear, non-technical, socially meaningful way

Key Concept 3: Be able to justify your outcome to affected stakeholders

(Leslie 2019, 35-36)

Conclusion: These three concepts again relate to stakeholders. The ideas behind them

are already implemented in Public Agency Manager Principles, Principle 4, and Vendor

Principles, Principle 4.

6.5.10 Democratic Legitimacy (Human Autonomy)

Key Concept 1: AI systems that “nudge” data subjects without their knowledge/consent might infringe on respect for human autonomy

(Leslie 2019, 5)

Conclusion: This concept is already full covered in Human Interaction Principles, Principle 7.

6.5.11 Democratic Legitimacy (SDPR)

Key Concept 1: The SUM Values

The “SUM Values” were devised by the ATI study as a combination of previous work in bioethics and human rights (Leslie 2019, 9-11). These values include:

- RESPECT the dignity of individual persons
- CONNECT with each other sincerely, openly, and inclusively
- CARE for the wellbeing of each and all
- PROTECT the priorities of social values, justice, and the public interest

More details for each of these values can be seen in their study.

Conclusion: While these values are undoubtedly important, as the ATI study’s author notes, they are “not specifically catered to the actual processes involved in developing and deploying AI systems” (Leslie 2019, 11). Rather, they are for helping to conceptualize what ethics and morals would make for good AI principles. Rather than the “SUM Values”, this study utilizes the idea of US substantive due process rights and subjective rights from the UN Universal Declaration of Rights. No new information is needed to be added.

6.5.12 Democratic Legitimacy (Accountability)

Key Concept 1: Automated decisions are not self-justifiable. Humans can be called to account for judgments, whereas machines cannot. This creates an accountability gap that must be addressed.

Key Concept 2: Automated decisions can be particularly complex in how they are brought about due to opaque black box models. This adds an extra layer of responsibility on a public agency to mitigate

Key Concept 3: Accountability can be broken down into answerability and auditability.

Key Concept 4: In terms of timeframe for implementation, accountability encompasses both anticipatory accountability and remedial accountability

(Leslie 2019, 23-26)

Conclusion: These four key concepts fit together, and they provide a compelling picture by the ATI study of what makes up accountability for public agencies. They will be integrated together into a new principle focusing on accountability under Public Agency Manager Principles.

6.5.13 Democratic Legitimacy (Deliberation)

Key Concept 1: Create a Fairness Position Statement (FPS) reviewable by all affected stakeholders

(Leslie 2019, 20)

Conclusion: Rather than focusing explicitly on algorithmic fairness as Optimization Principles, Principle 3 does, this concept takes a broader approach and includes non-algorithmic fairness as well. While Public Agency Manager Principles, Principle 1 mentions this idea briefly, it will be expanded upon. The idea behind an FPS, that a public agency should make clear to the public how it's defining fairness, is certainly important to make explicit.

Key Concept 2: Develop a Stakeholder Impact Assessment (SIA)
(Leslie 2019, 26-30)

Conclusion: This concept spells out Public Agency Manager Principles, Principle 4 in more actionable language. It will be used to update that principle and linked back to ATI's study for a more thorough review of what should go into an SIA.

6.5.14 Democratic Legitimacy (Maintainability)

Key Concept 1: Model accuracy can change as time passes and society itself shifts.
(Leslie 2019, 15)

Conclusion: This concept is covered in General Sociotechnical Principles, Principle 4.

Key Concept 2: Create a Dataset Factsheet
(Leslie 2019, 15-16)

The Dataset Factsheet in the ATI Study is just as the name implies – it's a factsheet for maintaining total data provenance, including issues such as "...procurement, pre-processing, lineage, storage, and security", plus qualitative input from team members regarding "...data representativeness, data sufficiency, source integrity, data timeliness, data relevance, training/testing/validating splits, and unforeseen data issues encountered across the workflow" (Leslie 2019, 15-16).

Conclusion: This concept will be used to expand upon the ideas presented in Vendor Principles, Principle 3. The original idea of a model factsheet will remain, but it will now be augmented to include the specific kinds of model-related data that it should

encompass. The principle's category location will also be changed to Public Agency

Manager Principles.

6.5.15 Democratic Legitimacy (Interpretability)

Key Concept 1: Logic, semantics, social understanding of practices/beliefs/intentions, and moral justification all play a key role in determining interpretability (Leslie 2019, 40)

The ATI study essentially takes *Section 4.2.2 What Makes a Good Explanation?* from this study and expands upon it to encompass this study's idea of interpretability. While that section of the literature review defines a good explanation as the balance between interpretability and completeness, ATI's study broadens the focus.

First, the ATI study asserts that *logic* plays a key role in determining explainability. While this is undoubtedly true, this idea should be too self-explanatory for inclusion as a principle itself – almost any element playing a key role in any decision requires logic to be valid. Second, the ATI study considers semantics. This concept most closely relates to *Section 4.2.2* in the sense that choosing proper semantics is a key element of how that balance is struck.

Third, the ATI study considers the agency's practices, beliefs, and intentions. This concept expands upon the Model Fact Sheet presented in *Human Interaction Principles, Principle 3*. It will be updated to further emphasize this element. Finally, the ATI study considers moral justification. While moral justification is also undoubtedly vital,

alongside logic, this study considers it to be self-explanatory – any public agency policy or activity in any circumstance should be morally justifiable.

Conclusion: Among the four concepts brought up by the ATI study, the idea of a social understanding of practices, beliefs, and intentions adds the most value for this study's framework. It will be added to the analytical framework. A new principle will be added to Public Agency Manager Principles.

Key Concept 2: Draw on standard interpretable techniques when possible
(Leslie 2019, 45-46)

Conclusion: This concept is a slightly more pointed take on General Sociotechnical Principles, Principle 2. No additional information is needed to be added.

Key Concept 3: Look first to context, potential impact, and domain-specific need when determining the interpretability requirements of your project
(Leslie 2019, 44-45)

According to the ATI study, *context* revolves around the type of application the machine learning system is to be used for. The requirements for a machine vision system used in policing should naturally be substantially higher than a machine learning system used to recommend the proper form that a website user needs to fill out. Embedded in the concept of context are potential impact (for example, is it assessing a high-risk activity) and domain-specificity (what kind of task is the system trying to perform).

Conclusion: This concept effectively spells out several important elements within interpretability in the context of public agencies. It will be incorporated as a new principle under Public Agency Manager Principles and combined with Key Concept 1 above and Key Concept 4 below.

Key Concept 4: When utilizing ‘black box’ AI systems (i.e. ANNs and SVMs), formulate an “interpretability action plan”
(Leslie 2019, 46-56)

The “interpretability action plan” would be designed to ensure the system provides effective explanations for the systems’ decisions, behaviors, and problem-solving tasks. The action plan would involve three elements:

- Clear articulation of the explanatory strategies
- Explanation delivery strategy
- Detailed timeframe for evaluating progress

Conclusion: This concept adds actionability to developing interpretability, and the framework presented by the ATI study is concise and useful. Rather than “reinventing the wheel”, this study will suggest that users implement the ATI study’s interpretability action plan.

6.5.16 Additional Concepts

The ATI study also focuses on three areas that were not explicitly covered in this study initially: *sustainability, safety, and the development lifecycle.*

Sustainability

First, in the case of sustainability the ATI study notes that “[d]esigners and users of AI systems should remain aware that these technologies may have transformative and long-term effects on individuals and society” (Leslie 2019, 26).

Conclusion: Sustainability is highly related to algorithmic maintainability and is also covered within the stakeholder impact assessment concept. No additional principles are needed.

Safety

The ATI study defines safety as a combination of “accuracy, reliability, security, and robustness” (Leslie 2019, 30). Accuracy and robustness each already have their own research thread in this study. Reliability is defined as an AI system performing as it was intended; this can be thought of as being encompassed within a combination of robustness and algorithmic maintainability. Finally, security is defined as a combination of data integrity (avoiding malicious modification of training/testing data) with data confidentiality (no unauthorized access to personal information), which can be mapped to a combination of robustness and privacy in this study.

Conclusion: No additional principles are needed.

Machine Learning Development Lifecycle

The ATI study also utilizes chronological phases (highly related to the software development lifecycle) for where in the development process various tasks occur. The

structure is ostensibly meant for analyzing fairness, but it is nevertheless applicable to the process overall (Leslie 2019, 22). The chronology consists of five phases:

1. Problem Formulation
2. Data Extraction & Acquisition
3. Data Pre-Processing
4. Modeling, Testing, and Validation
5. Deploy, Monitor, and Reassess

However, as noted in *Section 6.5.2*, the ATI study is lacking in regards to dealing with external vendors. Because of the frequency with which external vendors are used for developing ML systems in public agencies in the US, and since none of the original five phases deal with the vendor, a new stage (Vendor Negotiations) will be added in between Problem Formulation and Data Extraction & Acquisition:

1. Problem Formulation
2. Vendor Negotiations
3. Data Extraction & Acquisition
4. Data Pre-Processing
5. Modeling, Testing, and Validation
6. Deploy, Monitor, and Reassess

Conclusion: With the updated sixth chronological phase, this development lifecycle can be added to the analytical framework. This will be implemented throughout the framework as a simple chronological phase identifier in a new row to each principle's summary table. It should also help to answer Critique #5 from the previous stage in the methodology by tying the principles together more closely and showing how they relate to one another. In cases where a principle applies to more than one phase of the development lifecycle, this will be specified as well.

6.5.17 An Example for Applicability

This issue is one that I found applied to both studies after comparing them: a lack of a direct real-world example for applicability. Both studies relied strongly on theoretical examples, but neither provided an actual trained ML model with actual real-world data to attempt to test these ideas, even in a hypothetical public policy situation. To attempt to mitigate this limitation in my own study, I found a real-world public policy situation in which creating a ML model might be considered. While I could have used a ML system presently in use in a real-world situation, the lack of information about how these systems are used means that doing so would be problematic.

For this (admittedly hypothetical) public policy situation, I created several real ML models from real-world data in a situation where, if implemented, the ML models would have significant real-world consequences. More information on the ML models I created can be found in *Appendix B*, and will be cited throughout relevant areas of the final draft of my analytical framework below. In those cases where only a real-world case can be used (such as in the case of Vendor Principles – I needed no vendor), I will use the well-documented case of COMPAS as my example situation.

6.5.18 Summary of Key Improvements from Second Draft to Final Draft

The key improvements identified in this section going from the second draft to the final draft of the analytical framework are summarized below:

- Enhanced details of *General Sociotechnical Principles, Principle 3* to understand differing conceptions of fairness
- Enhanced conception of fairness from just fairness in outcome to encompass other areas of non-algorithmic fairness, including data fairness, design fairness, and implementation fairness
- Added additional question to *Public Agency Manager Principles, Principle 1*
- New principle with a more explicit focus on the role of accountability in democratic legitimacy is added to the framework
- Added additional language to enhance the ideas of making the public agency's position on conflicting fairness definitions known as well as how the public agency handles relevant stakeholders
- Expanded *Vendor Principles, Principle 3* to encompass the ideas from ATI's Dataset Factsheet
- Added new principle to Public Agency Manager principles which specifically focused on interpretability
- Added chronological information as to the development stage a given principle was most relevant in

6.5.19 Conclusion

In summation, the Alan Turing Institute's study on machine learning in public agencies is an excellent resource for a comparative analysis. While many of its key concepts were used to further refine and enhance this framework's existing principles,

its Stakeholder Impact Assessment and Interpretability Action Plan are arguably its greatest contribution to this analytical framework. These two concepts are clear, specific, well-reasoned, and do not try to provide all the answers. Rather, they attempt to guide the creation of the proper questions to ask, just as this study attempts to do.

6.6 Final Analytical Framework

The final product of my research methodology is the analytical framework below. Its principles come from a wide range of sources: some were found through archival research and my literature review, others came from peer review or expert interviews, and still more from my comparative analysis with the ATI study.

Each of the principles below is not designed to be a final answer or to argue in favor of or against a policy. Rather, they seek to provoke the questions that should be asked prior to and during the development and implementation of machine learning systems in public agencies. Those questions will have different answers in different use cases depending on both the normative values being applied to determine the “correct” answer as well as how technical techniques evolve and change. However, the principles behind them in this framework should remain more fixed. Some of the principles explicitly ask follow-on questions, whereas others allow the reader to determine their own follow-on questions as needed.

The framework’s principles are divided into five categories: (1) public agency manager principles, (2) general sociotechnical principles, (3) human interaction

principles, (4) optimization principles, and (5) vendor principles. These categories are determined from a combination of natural division points that appeared when extracting principles, as well as through expert interviews. The categories themselves are more fully defined below. While not all principles are necessarily phrased normatively for the sake of clarity and avoiding excessive repetition (for example, they don't all begin with "a public agency should make sure to consider..."), this preface can be presumed in those cases where it is needed.

At the beginning of each principle, I provide a summary table with the following characteristics for each principle:

Responsible Actor(s): Refers to the individual(s) within a given public agency that are likely to be the most responsible for implementing a given principle, if applicable. The options are Manager, Analyst, and Developer; while most public agencies have far more varieties of employees of course, these are generalized categories meant to provide broad guidance only. Brief descriptions for each role follow:

- *Manager* refers to the public agency manager who is responsible for the development of a particular ANN/ML system
- *Analyst* refers to the actual "low level" day-to-day internal users of the system within the public agency
- *Developer* refers to those internal agency software developers who are working with the vendor's software developers which are creating the ML system

Relevant Research Thread(s): Which research thread(s) are most relevant to the principle. In the case of democratic legitimacy, the *activity* within democratic legitimacy is specified as well.

Secondary Category (if applicable): While some principles fit cleanly into a single category, other principles have one primary category along with a second category that is also applicable, if not always quite as strongly. In cases where there is no secondary category, “N/A” is used.

Non-ANN Machine Learning Applicability: This identifies how applicable the principle is to other advanced ML systems such as random forests, support vector machines, etc. Some principles are fully applicable to both ANNs and other ML systems, others are only partially applicable, and some are simply only applicable to ANNs. This is either set to None, Partial, or Full. In the case of Partial especially, a short explanation is generally provided in the text of the principle itself.

Development Stage(s): This refers to the six chronological stages of development and implementation found in *Section 6.5.16* above.

Applied ML Model: This identifies which ML model I apply the principle to. The value will either be “DOHA”, “COMPAS”, “Both”, or “None”; None will be chosen when I do not believe there is value to be gained from applying the principle it after the fact. Both will be chosen when I believe that there is significant value in comparing how a principle would be applied to the DOHA model versus the COMPAS model.

The *DOHA* model refers to the artificial neural network I created in *Appendix B* from the Defense Office of Hearings and Appeal's (DOHA) publicly released data. The *COMPAS* model refers to the ML model used in the Wisconsin judicial system for recidivism prediction. It has been discussed several times in this dissertation previously, particularly in *Section 4.3* of the literature review.

For each principle, I determine which of the two ML models is the best to apply. For example, the DOHA Model is an artificial neural network and the COMPAS model very likely is not (we don't actually know precisely what kind of ML model it is due to vendor secrecy), which means that there are certain principles that aren't necessarily applicable to the COMPAS model. Likewise, the COMPAS model has actually been applied in the real world and the DOHA model has not, which means there are certain principles for which it is more applicable. Whichever model is chosen, I show how the given principle might shape the further development of the chosen ML model and/or whether the conclusions drawn from the principle might persuade or dissuade usage of the model at all.

Additionally, there are several terms that require further definitions for the following section:

Vendor - An external private firm which is responsible for the technical development of a machine learning system for a public agency.

Internal User - Public agency employees (usually more junior individuals in the agency's hierarchy) who are responsible for the day-to-day usage of the machine learning system within the agency.

External User - Members of the public who interact with and/or are assessed in some manner by the machine learning system.

Protected Groups - Groups which under US law are explicitly protected against discrimination. Includes subgroupings based on an individual's race, religion, national origin, age, and sex, among others.

6.6.1 Public Agency Manager Principles

Public agency manager principles are defined as principles that are oriented towards management of a machine learning project. They generally relate to issues occurring during problem formulation, or to issues that exist throughout the development process.

Principle 1: Definitions for key terms and concepts should be continually clarified and re-justified as development and implementation proceeds

Responsible Actor(s):	Manager
Relevant Research Thread(s):	All
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	All
Applied ML Model:	COMPAS

While this principle may seem self-explanatory, it bears further discussion. The myriad of taxonomies presented in this study, not to mention those presented in others, should show both the incredible breadth of techniques that exist and the amount of critical terms with subjective and debatable definitions. How you define (and how you justify your definitions) for the critical concepts comprised within each research thread will play a substantial role in just about every successive activity. Given the complexity and subjectivity of several of the key issues, this is not an area to be overlooked. What's

more, definitions are not static – they can and should be refined as development progresses. Indeed, such refinement may be required since not all information may be known during the *Problem Formulation* stage.

In the case of COMPAS, we could certainly see a significant potential improvement from applying this principle from early on. For example, the firm that produced COMPAS, Northpointe, was permitted to define fairness themselves; there was no public or agency input into that decision. Because of that, it has come under frequent critique for not having a sufficiently strong algorithmic definition of fairness (see *Section 4.3*). Were such a determination to be made at the public agency level and continually re-justified based on public feedback, it would be much more difficult to attack in such a way.

Several follow-on questions arise from this principle:

Question 1: How much explainability is enough for your machine learning system and why?

In the case of COMPAS, I argue that a pretty substantial amount of explainability should be required before usage. At the very least, there should be additional information as to how COMPAS' recidivism likelihood statistic was calculated. This means sharing the particular ML algorithm used to create COMPAS. Depending on the particular algorithm used, more specific explanatory information would change.

Question 2: What type(s) of privacy are you implementing (differential privacy, federated learning, etc.) and why?

To my knowledge, no particular privacy protocols were implemented on COMPAS algorithmically. However, since the overall threat of de-anonymization is fairly low, this likely isn't a significant problem. The dataset Northpointe released is fully anonymized, with almost nothing in the way of identifying an individual's characteristics with their real-world identities. Standard cybersecurity practices to protect the information should be sufficient.

Question 3: Does the agency consider fairness in the algorithm's design, fairness in the model's outcomes, fairness in data collection, and fairness in implementation (i.e. effective internal user training)? How are each of these definitions justified?

Unfortunately, this question is a big question mark for the case of COMPAS. We don't know precisely how fairness was incorporated into data collection, implementation, outcomes, or design. This is a major issue with how COMPAS' proprietary nature is harmful.

Question 4: How were adversarial examples and data poisoning mitigated?

Again, we don't know how these issues were mitigated, if they were at all. However, the risks in both areas are likely to be rather small with COMPAS – there is little opportunity for adversarial examples, and data poisoning is also minimal unless the datasets were hacked into.

Principle 2: The importance of each research thread should vary between different types of machine learning and different use cases

Responsible Actor(s):	Manager
Relevant Research Thread(s):	All
Secondary Category (if applicable):	N/A

Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	All
Applied ML Model:	Both

Permanently “ranking” research threads (save for democratic legitimacy as the most important) is a poor idea for a public agency due to the changing importance of each thread from case to case. For example, if there is no interface through which malicious input could be provided (as in the case of COMPAS, where prisoners’ data is not entered in by the prisoners themselves), optimizing for robustness becomes less important. Likewise, if the data involved isn’t personally identifiable information and is already available to the public, the value of optimizing for privacy is also reduced. That’s not to say that they are entirely unimportant in those instances, but simply that considering other research threads should be of greater importance.

For the COMPAS model, I argue that the importance of each thread ranks as such:

1. Democratic Legitimacy
2. Fairness
3. Explainability
4. Accuracy
5. Privacy
6. Robustness

In contrast, for the DOHA model, I argue that the importance of each thread ranks as such:

1. Democratic Legitimacy
2. Accuracy/Fairness
3. Explainability
4. Privacy
5. Robustness

Both lists are similar, but there are some distinct differences as well. Democratic legitimacy is naturally on top for both cases, and in both cases the fear of adversarial attack is pretty minimal. For the DOHA model, it's difficult to determine whether or not accuracy or fairness are of greater importance – in the real world I would want to go back to DOHA and see if they had additional personal information I could use to assess fairness and bias for issues such as race, ethnicity, etc. I would also want to see how trying to maximize fairness could potentially damage predictive accuracy.

In the case of COMPAS, accuracy is hardly unimportant but the needs of fairness and explainability are potentially higher. Particularly in the context of our judiciary, it is exceptionally important that we be able to explain why certain decisions were made. In addition, the US has historically had unfairness towards protected groups in our justice system, requiring that issues of fairness be taken to the forefront. However, if ensuring that there is sufficient explainability and fairness causes accuracy to degrade too much, this may indicate that a machine learning system is unsuitable for the task at hand.

By contrast, explainability may not be of the highest importance for the DOHA model since isn't the true decision-making entity (it's just reading the text of the case summary; see *Appendix B*), but it is still having enough of an impact that it is reasonable to want to understand why the DOHA model makes the decisions it does.

Principle 3: Public agency decision-makers should be prepared to engage in quantitative coding of value judgments that were previously made qualitatively

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Interpretability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Data Pre-Processing
Applied ML Model:	None

Public agencies often have qualitative values encoded into the decisions that the agency is responsible for. These qualitative values are not always clear-cut and can have a significant amount of subjective value assessment. One possible problem when trying to implement machine learning systems for such cases is that these systems cannot use these purely qualitative (and often intuitive) value assessments as inputs – it requires those inputs to be quantified to some degree. This change from qualitative to quantitative values may be difficult for some agencies to manage and should be scrutinized during development and implementation.

It is particularly difficult to apply this principle to either model. It cannot easily be applied to the DOHA model because DOHA is not used in the real world, and it cannot easily be applied to the COMPAS model because we do not know enough about the internal non-judicial decision-making in the Wisconsin judicial system.

Principle 4: Accountability should not just be remedial, but anticipatory as well

Responsible Actor(s):	Manager
------------------------------	---------

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Interpretability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Deploy, Monitor, and Reassess
Applied ML Model:	COMPAS

Accountability is not something just done after the fact – there are anticipatory activities a public agency can do to help mitigate future accountability issues when utilizing ML systems. These accountability activities can be divided into procedures to increase answerability and procedures to increase auditability. *Answerability* means ensuring that there is a continuous chain of human responsibility across the entire workflow, with those responsible able to provide at least some level of explainability, while *auditability* means “demonstrating both the responsibility of design and use practices and the justifiability of outcomes” (Leslie 2019, 24).

Whereas answerability and auditability involve the tasks that need to be done to make up accountability, however, anticipatory and remedial accountability focus on the *when* the accountability should take place. Anticipatory accountability involves ensuring that there is accountability by design throughout the design and implementation process of an AI system, and remedial accountability involves the processes setup to deal with problems in accountability that occur after the fact.

Together, these ideas produce the following matrix:

Table 20 - Answerability vs Auditability

	Answerability	Auditability
Anticipatory accountability	There should be a continuous chain of human accountability during the design and implementation of the ML model.	All system design decisions should be fully justifiable during the design and implementation of the ML model
Remedial accountability	There should be a continuous chain of human accountability during the remediation of accountability issues discovered after implementation.	All system outcomes should be fully justifiable during the design and implementation of the ML model

A further review of these ideas can be seen in the ATI study (Leslie 2019, 24).

For the COMPAS model, none of the boxes in the matrix above have a clear answer: we simply don't know who is accountable for COMPAS' outcomes. It might be argued that Northpointe itself is the answer to all four aspects of accountability, but accepting that explanation adds additional problems - should a private company really be the one accountable for (one element of) the decisions of Wisconsin's judicial system?

Principle 5: Utilize a Model Factsheet

Responsible Actor(s):	Manager Developer
Relevant Research Thread(s):	Explainability, Democratic Legitimacy (Accountability, Transparency, Interpretability)
Secondary Category (if applicable):	General Sociotechnical Principles
Non-ANN Machine Learning Applicability:	Partial

Development Stage(s):	Data Pre-Processing Modeling, Testing, and Validation Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

Some ML models (particularly ANNs) are not very forthcoming about what makes them tick. The output itself is clear, but the reasoning behind that output is often much more difficult to parse. While there is a limit to what the model itself can explain, however, maximizing interpretability is reliant on the public agency itself. Even if state-of-the-art explanatory techniques are successfully applied to a given ANN, these explanations will be of little value if the public agency does a poor job of providing socially meaningful content to the public. For both explanations of the meaningful content a model produces and for more structural information regarding how the model was created, a model factsheet can be a useful solution. In short, this model factsheet is one of the essential mechanisms through which explanatory power and interpretability can be maximized – a simple yet powerful method for explaining a model and its data.

A model fact sheet is a standardized, relatively non-technical outline of the capabilities and limitations of a given ANN, although it could be applied to non-ANN machine learning systems as well. Depending on the use case, the precise content can vary greatly. (Brajer, et al. 2019) provide an example of a model fact sheet in the case of healthcare delivery:

Model Facts
Model name: In-hospital death risk prediction model
Summary This model uses EHR input data collected between a patient's time of presentation to the hospital and admission to the hospital to estimate the probability that the patient will die during the hospital stay. It was developed in 2018-2019 by the Duke Institute for Health Innovation.
Mechanism • Outcomein-hospital death • Output0%-100% probability of death occurring during the hospital admission • Patient populationall adults >18 years old admitted to the hospital • Time of predictionat the time of admission to the hospital • Input data typeEHR • Input data sourcedemographics, labs, vitals, medication administrations • Training data location and time periodhospital A, 2014-2015 • Model typeXGBoost
Validation and Performance • Retrospective: tested in 2018 at 3 hospitals (A, B, and C); AUROCs 0.84-0.89, AUPRCs 0.13-0.29 • Prospective: tested in 2019 at one hospital (A). AUROC 0.86 and AUPRC 0.14. • Operational: at hospital A, with threshold set to achieve 20% PPV, 51% sensitivity, and 96% specificity, for 100 admissions per day, there will be approximately 6 total alerts, 5 false alerts, and 1 true alert.
Uses and directions • General use: this model is intended to be used as an additional source of information for clinicians making operational and clinical decisions while caring for a newly admitted hospital patient. Specifically, this model is intended to be used to identify patients at high risk of in-hospital death • Tested use case(s): • Examples of appropriate decisions to support: identification of high-risk patients for consideration for services that they would have received later or not received in the absence of information provided by the model • Before using this model: test the model prospectively on the local data and confirm the ability to monitor model performance over time.
Warnings • General warnings: this model may take expected treatment effects into account. The degree to which the model anticipates treatment has not been studied and is unknown. An example of this occurring would be a young patient with a life-threatening, traumatic wound being categorized as low-risk because, with standard treatment, the patient would have a low risk of dying in the hospital. Therefore, the model should not be used to triage patients for the ICU. • Examples of inappropriate decisions to support: deprioritization of low-risk patients for services that they would have otherwise received with standard, usual care, in the absence of information provided by the model. • Discontinue use if: clinical staff raise concerns about how the model is being used or model performance deteriorates owing to data shifts or population changes.
Other information: • Publications: -- • Related models: --

Figure 13 - Model Factsheet Example

Credit: (Brajer, et al. 2019)

We can see several critical areas from *Figure 13* above:

- Name of the model
- Mechanisms for input, outcome, data, and model information
- How it was tested and should be used

- Warnings about potential inappropriate uses and the conditions under which it should be discontinued
- Other models that are related to it

Regardless of the use case, such a fact sheet should allow non-expert users to both understand how a model works and allow for easier model-to-model comparisons and analysis. The information itself can be provided through any or all of the explanatory techniques discussed in the literature review, such as processing techniques, representation techniques, attention techniques, etc.

Additionally, such a factsheet should cover issues of data and model provenance as well – where did the data come from and how has it been transformed since then. ML systems can have data arrive from a wide array of sources. That data can be split, rearranged, transformed, merged, reconstructed, and extrapolated multiple times from initial data ingest to training. The determination of where this data came from, whether it is accurate, and how it is used are questions of significant importance when public agencies implement machine learning systems (Shrum, et al. 2019, 19). Interpretation of the output of a ML system requires (to some degree) the ability to determine where the data came from and how valid that data is. Likewise, external reviewers require data provenance to ensure public agency accountability.

It also may be necessary to create a separate model factsheet for both internal and external users. This is due to several reasons. First, the data that both groups are interested in may vary. Internal users may be more interested in the granular details of the meaning of a model's output. By contrast, external users may be more interested in

information relating to how the particular model was chosen and what intentions and beliefs went into creating the model in the first place.

Additionally, not all information may be able to be released to the public as compared to internal users. In a perfect world, the public would be able to see the entirety of the data used to train the model, as well as the precise details of the model's structure. However, issues of trade secrecy (see *Vendor Principles #1*), privacy and robustness (see *Optimization Principles #4*), and de-anonymization (see *Optimization Principles #5*) make such absolute transparency non-viable.

For the DOHA model, below is what a Model Factsheet might look like:

Model Facts
Model Name: DOHA Model Summary: This model uses textual clearance case summary data gathered from investigators involved in the determination of whether potential Department of Defense contractors should be granted security clearances. When someone is initially denied a security clearance, they can appeal their case to the Defense Office of Hearings and Appeals (DOHA). This dataset uses those first-level appeal cases; it ignores second-level appeals.
Mechanisms - Outcome: Security clearance granted or denied - Output: Probability from 0%-100% on whether the clearance should be granted - Data Population: all defense contractors from the DOHA website at the first stage of their appeal process; about 5,000 clearance case summaries in total - Time of Prediction: the date their clearance case summary is archived publicly on DOHA's website - Input Data Type: Text - Input Data Source: DOHA website - Training Data Time Period: 1997-2019 - Model Type: Convolutional Neural Network
Validation and Performance Recall: 0.957 Precision: 0.966 F1 Score: 0.962 Informedness: 0.9411 Markedness: 0.946
Use and Directions - General Use: this model is intended to be used as an additional assessment check for administrative judges at the Defense Office of Hearings and Appeals.
Warnings - This model is based on human-written textual assessments of individuals. Because of this, changes in such elements as writing style, content, or clearance approval standards may cause this neural network to become substantially less accurate.

Figure 14 – DOHA Model Factsheet

Principle 6: Determine the relevant stakeholders and invite stakeholder feedback

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Deliberation, Transparency)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

Public review of new regulations or policies is a frequent feature of many public agencies, such as the Federal Communications Commission (FCC 2019). However, such reviews become increasingly important when machine learning systems come into play. This is because even with public or expert input, there can be a significant lack of explainability in the results of a machine learning system.

More specifically, the ATI study provides documentation for how to conduct what it refers to as a *Stakeholder Impact Assessment*. This assessment is designed to build public confidence in the design and deployment of the ML system, strengthen accountability, bring to light unseen risks, enhance transparency, and demonstrate a public agency is doing their due diligence to the public. Rather than repeat them verbatim here, more details on the specifics of this assessment can be seen in their study (Leslie 2019, 26-30).

Relevant stakeholders for the DOHA model (aside from a general public interest) include the Department of Defense Inspector General, Congress, DOHA's own

administrative judges, and applicants to the federal government who might need security clearances.

Principle 7: Public agencies should develop protocols early to maximize interpretability

Responsible Actor(s):	Manager Developer
Relevant Research Thread(s):	Democratic Legitimacy (Interpretability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Problem Formulation Vendor Negotiations Deploy, Monitor, and Reassess
Applied ML Model:	None

While all ML systems need interpretability, that interpretability can be more difficult to come by ANNs especially. Furthermore, interpretability is not an afterthought to be considered at the end of development. Rather, it is essential that it be considered during initial problem formulation and vendor negotiations. This is because it can require early intervention in an ML system's development to ensure it has sufficient interpretability.

The ATI study recommends what it calls an *Interpretability Action Plan* to maximize interpretability. While their study is not the only one that considers what interpretability in ML should entail, it is among the only studies that considers interpretability specifically from the lens of public agencies. For example, (Lipton 2016) provides a far more exhaustive review of what interpretability in machine learning

should entail generally. For precise details on how to construct an Interpretability Action Plan, the ATI study can be reviewed (Leslie 2019, 46-56).

6.6.2 General Sociotechnical Principles

This category includes broad principles focused on where the more technical aspects of machine learning system development are subjective and require more than just a degree in computer science to deal with them effectively.

Principle 1: Correlation is not causation

Responsible Actor(s):	N/A
Relevant Research Thread(s):	Fairness, Democratic Legitimacy (SDPR)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Modeling, Testing, and Validation
Applied ML Model:	DOHA

The statistical adage “correlation does not imply causation” is well-known (S. Singh 2018). Admittedly, this principle is not actually a normative principle like the others, but rather a positive one; this makes it rather unique in this analytical framework. Nevertheless, it has such a special and important meaning in the context of machine learning systems and ANNs, particularly as applied to questions of public policy, that I believe it requires its own principle regardless.

Simply put, machine learning systems do not predict causation; rather, their predictions are only based upon countless subtle correlations. The output should thus never be thought of *on its own* determining a direct causation between the input and

the output. This is particularly important considering that the end-users of a given ML system, whether it be the public or the average analyst at a public agency, may be unlikely to be well-versed in statistical analysis (Faes, et al. 2019).

Examples abound on the gap between correlation and causation. Consider the following examples from the book/website *Spurious Correlations* (Vigen 2015):

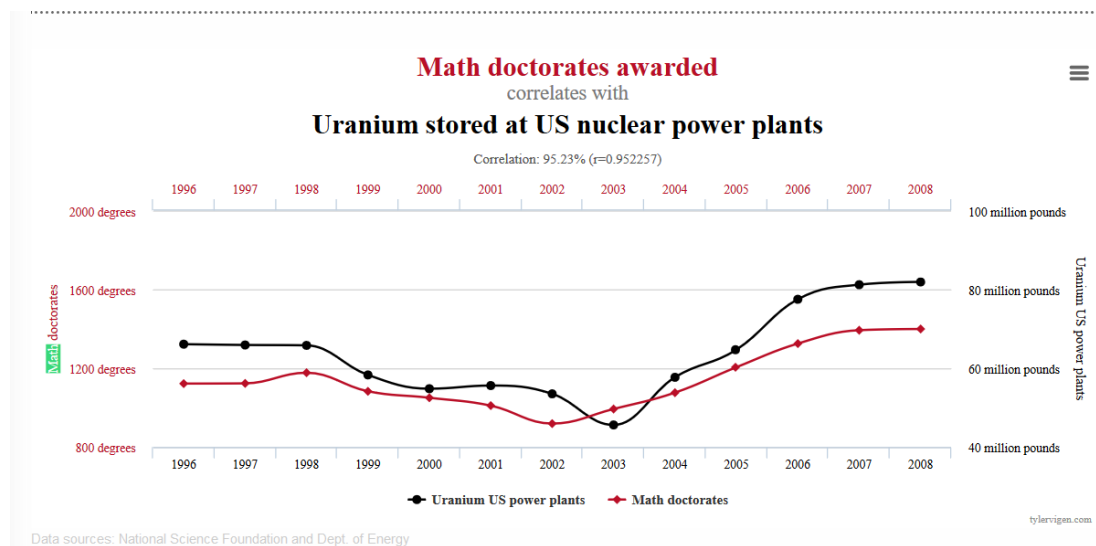
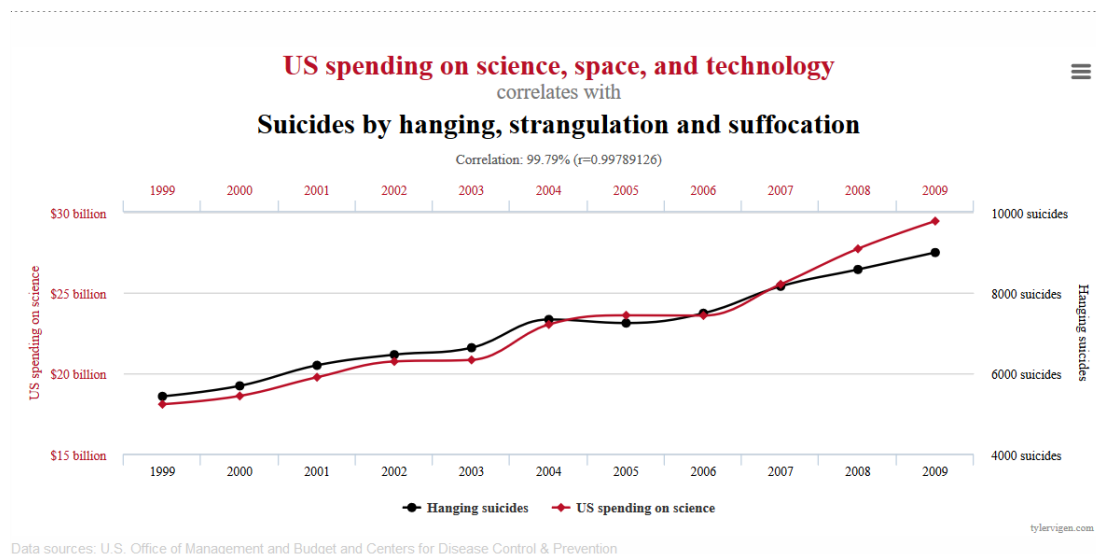


Figure 15 - Correlation != Causation Examples

Obviously, there is no *causal* relationship between math doctorates and uranium storage or U.S. science R&D spending and suicide rates, yet both relationships have a correlation of over 95%. What can this tell us about machine learning systems? Simple - if you provided an ANN with the math doctorates awarded each year as input and had it attempt to predict the uranium stored at US nuclear power plants as output, it would likely perform admirably well. However, the deeper question underlying this for public agencies to deal with is whether making determinations based upon a given correlation is a positive or a negative – should two variables unlikely to have a causal relationship be used to predict one another purely based on a correlation?

Question 1: Should a causal relationship between input and output be required (or at least asserted) prior to using a given variable as the input?

Question 2: What is the standard for determining causality in such cases?

Question 3: Are there procedures in place to try and test which variables are correlated, even in black box systems?

Answering all three questions for the DOHA model, there should be a clear causal relationship to the model's data: what the administrative judge writes about a given applicant's case should cause the applicant to be granted or denied a security clearance. However, as previously noted it is also possible that the model is focusing on seemingly unimportant but correlated text within each clearance case summary.

Convolutional neural networks like the DOHA model have several techniques (some discussed in the literature review) to help understand what elements of the input the neural network is most focused on – this should help to answer the second question.

Were this model to be implemented, DOHA would need to create procedures to test for this.

Principle 2: Public agencies should test multiple types of machine learning systems

Responsible Actor(s):	Developer
Relevant Research Thread(s):	Fairness, Accuracy, Explainability, Democratic Legitimacy (Interpretability, Transparency)
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Modeling, Testing, and Validation
Applied ML Model:	DOHA

Even if ANNs have achieved broadly superior results to more classical ML techniques in the realm of raw accuracy, the potential loss in terms of other research threads may indicate that an ANN is not the best solution for every case. What’s more, there are still some areas where other machine learning systems can still match an ANN’s results in terms of raw accuracy. Thus, rather than settling on ANNs from the beginning in order to use the “most advanced” technology, it is worthwhile to have multiple kinds of systems built and tested. Different machine learning systems suffer from different flaws which may be more or less important in different use cases.

For the DOHA model, this was certainly conducted. Four other types of ML models were tested alongside the convolutional neural network (see *Section B.5.2*) to see how the DOHA model compared. The DOHA model surpassed all of them in predictive accuracy as ANNs often do, although the other methods were all less of a

black box than an ANN is (except perhaps for the SVM). The question, then, of what percentage decrease in accuracy is worth what level of increase in explanatory power, is the vital one to answer.

Principle 3: Unrepresentative data should be checked for even in cases where an entire population is the sample

Responsible Actor(s):	Developer
Relevant Research Thread(s):	Accuracy, Fairness, Accuracy, Democratic Legitimacy (non-algorithmic Fairness)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Data Extraction & Acquisition Data Pre-Processing
Applied ML Model:	Both

I use the term “unrepresentative data” for this section in contrast to “sampling bias” because an entire group or population can *be* the dataset when it comes to training machine learning systems, yet still have fundamental problems related to the sample obtained. Under the traditional sampling bias paradigm, no sampling means no sampling bias. However, if your data isn’t representative of reality, either due to how it was collected or social policies which affect who is in the dataset, insights gleaned from it that were trained on that data may be tainted and produce unfair results. This is almost regardless of the model’s predictive accuracy.

For example, going back to the case of COMPAS: if it were the case that US policing policies caused unrepresentative populations of protected groups from the

overall US population being arrested and imprisoned to be used as training data to predict recidivism rates, these biases could be transferred into a model.

Because of this, declaring predictive accuracy or data labels as the equivalent of ‘ground truth’ may *itself* be flawed in some instances. Rather, accuracy may more closely resemble the idea of being *consistent* with previous decisions. An ML system trained on data that was labelled by human subject matter experts (SMEs) thus is not necessarily predicting accuracy to reality, but rather predicting what an amalgamation of SME analysis would predict (which we hope is equivalent to reality). This relates to the ideas discussed in fairness literature relating to the *observed space* versus the *constructed space* (see Section 4.3.2).

For its own part, the DOHA model could easily have unrepresentative data. Even though there was no sampling (to my knowledge), the DOHA data may not include all records from 1998 onward (perhaps DOHA doesn’t put every record online for access, for example). Additionally, perhaps DOHA’s standards have changed over time from 1998 to 2019; if these clearance case summaries were written substantially differently in the past or even by the different styles of writing by different judges, this could potentially make the data substantially unrepresentative. Finally, DOHA’s data only included DoD contractor applicants and thus it may not be fully applicable to non-contractors seeking clearances if significant differences in those populations were to be observed.

Principle 4: Procedures for “algorithmic maintenance” should be defined from the beginning of development

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Accuracy, Democratic Legitimacy (Algorithmic Maintenance, Accountability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Problem Formulation Vendor Negotiations Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

Algorithmic maintainability (as applied to ANNs in particular, though somewhat applicable to ML systems generally) is the idea that there should be specific procedures in place to handle how these systems should be periodically analyzed and updated to ensure that they are maintaining the same level of performance as when they were initially assessed. Unlike traditional software packages, updating a machine learning system automatically just for the sake of updating it may not be the correct solution. This is because unlike traditional software packages, a “new” version should not be construed as automatically implying the system is inherently superior. Rather, what that update would entail is of great importance: often an update entails additional training data being used to try and improve the model. Such updates need to be considered on a case-by-case basis.

Several questions arise from this principle:

Question 1: Should new data be inputted continuously into an ANN's training as it becomes available, or should it be done with "versions" similar to traditional software development methods?

For the DOHA model, there would need to be a deep review of how predictive accuracy changes when new clearance case summaries are provided to the model.

Question 2: Should the ANN be retrained every N years from scratch?

I believe that the DOHA model should be retrained from scratch every N years (the precise number would require deeper study) because judges change and even the standards of assessment may also change over time. Because of this, older records may not be as valuable as newer records if standards and styles of writing change over time. However, this retraining is predicated on the model's accuracy not degrading from fewer training samples. If the model's accuracy would decrease, this makes the question much more complex to answer, and it becomes a balancing test.

Question 3: Who is responsible for the decisions stemming from the ANN?

DOHA would need to assign an individual or a specific team of individuals responsibility for these algorithmic maintenance questions. It would be their responsibility to answer these questions in greater depth and report on their findings whenever an update is considered.

Principle 5: Public agencies should measure more than just the raw "accuracy" statistic for a given ML model

Responsible Actor(s):	All
Relevant Research Thread(s):	Accuracy
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Modeling, Testing, and Validation

Applied ML Model:	DOHA
--------------------------	------

As the *Accuracy* section in the literature review should have made clear, even the study of accuracy itself (generally thought of as the most quantitative and straightforward of research threads) can still have qualitative subjectivity. Public agency managers and internal users both need to be aware of what kind of “accuracy” numbers a ML system is providing, or if the system is assessing on a unique algorithm such as the BLEU score. The difference between assessing actual raw accuracy and F-1 Score can be substantial when considering the suitability of a ML system for implementation.

The DOHA model measures not only raw accuracy, but also recall, precision, F1 score, markedness, and informedness. All of these statistics produced very strong results. Doing so ensured that the accuracy statistic on its own was not hiding deeper problems in the model.

Principle 6: Protecting underrepresented groups within training datasets should be of particular importance since they are often in greater danger from poor model performance

Responsible Actor(s):	All
Relevant Research Thread(s):	Democratic Legitimacy (SDPR), Accuracy, Fairness
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Data Extraction & Acquisition Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

Protected groups can be intrinsically at greater risk of unfair treatment by a public agency because there often won't be as much training data for the minority group (Zhong 2018). The correlation between the quantity of high-quality training data and model performance is extremely strong; all other things being equal, a model with significantly more training data will almost always achieve equal or (more often) superior performance to an equivalent model with less training data. Because of this, protected groups may be particularly vulnerable to a dearth of training data. While there are various techniques which attempt to mitigate these issues, there is presently no "silver bullet" solution (Barros, et al. 2019).

Question 1: Does your training dataset have relevant protected groups? If so, are you testing the predictive accuracy for those groups separately?

Unfortunately, only the gender of applicants is known, not other protected statuses. If those protected statuses are known to DOHA officials, they should also be tested for potential bias in the model, particularly underrepresented groups. However, gender was tested and it definitively showed that accuracy did not degrade below 97% for either gender. No matter which standard of fairness is chosen, we can say that the model is extremely unlikely to have gender bias.

Question 2: Is your definition of algorithmic fairness mitigating this problem?

In the case of gender only, the DOHA model meets or nearly meets both parity and equality of odds, as well as the other standards suggested by the ATI study in *Section 6.5.5*. As stated previously, without other protected group information we do not know for certain since that information is not available.

Question 3: Are there likely to be new protected groups with specific unique characteristics that are presently missing from the training dataset?

Once again, this is unknown for the DOHA model – we do not know what protected groups we *do* have, so it’s hard to say which ones we don’t have too.

Principle 7: Public agency analysts should question assumptions about what they think is being learned

Responsible Actor(s):	Analyst
Relevant Research Thread(s):	Accuracy, Explainability, Democratic Legitimacy (Interpretability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Modeling, Testing, and Validation Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

Machine learning systems, and particularly ANNs, will often try to “cheat” during training. By design they seek the easiest training path to maximize accuracy. For image recognition tasks, this often (though not always) means that the most obvious and consistent differences between two different image classifications will be what the system learns. While this may seem like a good thing at first glance (and it often is), it isn’t always a positive.

Indeed, there are several famous cases in the history of machine learning where this misunderstanding of what was learned caused substantial real-world problems. One of the most well-known cases was when a neural network was trained to differentiate between wolves and huskies (Ribeiro, Singh and Guestrin 2016). On the one hand, the

model achieved a very high accuracy during training and it was thought of as a great success initially (particularly back in 2016).

However, when it was applied to the real world, it failed spectacularly, misclassifying what should have been easy identifications between wolves and huskies. When researchers dug into the ANN, they found out that they had been wrong about what it had succeeded in classifying: the ANN hadn't been classifying the animals, but rather it had become adept at identifying images with snow in them. Since all the images with wolves had snow in them, it was thus able to successfully classify the wolves rather easily. In the real world, however, the images of wolves did not always have snow in them. This caused the classifier to essentially malfunction (Kepler 2019).

For the DOHA model, what we think is being learned is that the DOHA model is picking up on key phrases (and variants of those phrases) that it has learned should generally cause someone to be granted or denied a security clearance, as well as the relationships between those key phrases. However, this assumption may not be correct. As noted previously, it is possible that the DOHA model is merely focusing on highly correlated phrases that nevertheless should not reasonably be related to whether someone should be granted a security clearance. In the real world, additional experimentation should be conducted to determine which kinds of words or phrases the DOHA model is focusing on most often.

Principle 8: Variable selection should not be left entirely to computer scientists

Responsible Actor(s):	Manager Developer
Relevant Research Thread(s):	Explainability, Democratic Legitimacy (non-algorithmic Fairness, Accountability)
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Data Extraction & Acquisition Data Pre-Processing Modeling, Testing, and Validation
Applied ML Model:	DOHA

While one of the greatest strengths of ANNs is their ability to find the most important variables themselves (not to mention the subtle relationships between those important relationships), this strength can also be a weakness. In most other kinds of machine learning, the human is generally responsible for selecting the most important variables, but this isn't always the case with ANNs. However, while the model chooses the relevant variables from among those provided to it, it is the responsibility of those developing and implementing the model to choose which variables the ANN can choose from in the first place. While obvious protected group indicators (depending on the use case) may require being excluded (i.e. race, sex, religion, etc.), more subtle variables may *implicitly* encode various protected statuses anyway (Roberts 2018). Thus, finding the proper balance between eliminating unwanted correlations and maximizing model accuracy should be a constant consideration.

It is thus unwarranted to leave such a decision to those responsible for developing the ANN alone. Rather, the usage of particular variables can have major implications for democratic legitimacy and explainability – public agencies cannot expect to retroactively justify using questionable variables by blaming the computer scientists.

For the DOHA model, there were several subjective decisions that were made about the data to be used. First, only the two sections of text which I deemed to be the most objective were used to avoid the DOHA model applying circular logic. Nevertheless, I could have been mistaken in my assessment – perhaps only one of those sections should have been included, or perhaps there was an additional section of text I missed which would have substantially improved the model.

Second, I also gathered binary True/False information on what formal suitability criteria an applicant had problems with. Since this dataset involved appeals cases, all individuals had at least one suitability criteria where the initial determination was made that the individual had a problem. However, since I achieved such a high level of accuracy simply by using the text of the clearance case summary, I did not use this suitability criteria information as additional input for my model. This decision may also need to be considered further.

6.6.3 Human Interaction Principles

There are two potential human “audiences” for an ML system implemented in a public agency: *internal users* and *external users*. These terms were defined at the beginning of this section. Some ML systems will only have internal users and some will have both internal and external users; it’s highly unlikely that an ML system will have no internal users at all in the context of public agencies and within the scoping of this study. Whether internal or external, the principles below focus on how humans interact with ML systems.

Principle 1: Public agencies should design the user interface (UI) as a critical feature, not an afterthought in machine learning system design

Responsible Actor(s):	All
Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Interpretability), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Vendor Negotiation Deploy, Monitor, and Reassess
Applied ML Model:	COMPAS

This principle is emphasized by (Pásztor 2018) most prominently: a computer system is not simply its input and output, it’s also its *interface*. All too often, government systems have terrible user interfaces (Sinders 2018), and that can cause significant issues in the development and usage of ML systems. These issues can be for internal users and external users both.

For example, a poor user interface for internal users may allow flaws and/or biases to go unnoticed. Likewise, a poor user interface for external users may provide an incorrect sense of what the system determined and why it determined it. While government has traditionally had a poor history of UI design, the problems arising from ML systems are not just frustrating but can have significant negative real-world consequences. Some recent scholarship has focused on experimentation with using interfaces to increase the understanding of trade-offs between accuracy and fairness (Yu, et al. 2019).

In the case of COMPAS, I don't know precisely how this recidivism information was provided to judges in terms of user interface. However, particularly prior to the case of *Wisconsin v. Loomis*, there was clearly little to no information for the judge to understand where this kind of statistic came from. One way of applying this principle would be to ensure that the information provided by COMPAS be colored differently than other information, and surrounded with key contextual information in a non-technical format about where the statistic comes from and what, precisely, it indicates beyond a simple "risk number."

Principle 2: Internal users should be enabled to do their own testing

Responsible Actor(s):	Analyst
Relevant Research Thread(s):	Democratic Legitimacy (Algorithmic Maintenance, Interpretability), Explainability
Secondary Category (if applicable):	General Sociotechnical Principles
Non-ANN Machine Learning Applicability:	Full

Development Stage(s):	Vendor Negotiations Deploy, Monitor, and Reassess
Applied ML Model:	COMPAS

The design of almost any ML system should allow for internal users to do their own testing and analysis. For example, depending on the sensitivity of the data and how “fakeable” it is, internal users should be able to enter fake data to help them get a better understanding of what the system is capable of and to spot potential design flaws. Even if internal users aren’t computer scientists, they’re more likely to be subject matter experts and may detect problems that would be otherwise missed.

Beyond simple testing, complex analytical suites are also available such as Stanford’s open-source neural network verification project (SyncedReview 2019). Such analytic suites can be important to help internal users grasp if an ML model is behaving as it should be even before external experts or consultants are hired to review it. In the case of COMPAS, it would be incredibly useful for employees in the Wisconsin judicial system to be able to send “fake” prisoner information into COMPAS to assess how its output changes. For example, out of the over 100 input variables that COMPAS accepts, internal users could test how substantially changing just a single variable might alter COMPAS’ output. This could also be used to help detect biases against characteristics common with protected groups.

Principle 3: Differing user acceptance of False Positives versus False Negatives should be assessed for each use case

Responsible Actor(s):	Manager Developer
Relevant Research Thread(s):	Democratic Legitimacy (Accountability), Explainability, Accuracy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Modeling, Testing, and Validation
Applied ML Model:	DOHA

Depending on the use case, internal users may be more or less receptive to false positives versus false negatives. In other words, there can be a significant difference between stating that a wrong answer is right versus stating that a right answer is wrong. This is entirely dependent upon the use case, however. For example, consider the case of an image classification system designed to track poachers (Harvard 2019). For those internal users assigned to understand what the system is saying, there can be a significant difference in user acceptability between showing too many false positives (that is, showing that poaching was occurring when it actually wasn't) and false negatives (showing that poaching was not occurring when it actually was). Even though the raw accuracy may be the same regardless, the internal users in this case are much more likely to be willing to get false positives rather than false negatives: better to sift through the false positives to find the real cases of poaching rather than miss actual cases of poaching entirely (within reason).

For the DOHA model, user acceptance likely tilts in favor of false negatives over false positives, although neither can be easily dismissed as unimportant. Predicting that an applicant should not be granted a security clearance when they should be is less than ideal since a qualified individual will not be able to do their job. However, predicting that an applicant *should* be granted a security clearance when they should not is a whole different issue in that it is a potential threat to national security. Thus, the model should be scrutinized closely for whether it is providing more false positives or false negatives. Even if it reduces accuracy, the model might still be improved overall if it used constrained optimization it better minimize false positives over false negatives.

Principle 4: Public agencies should explicitly determine what information should and should not be provided for external users

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency), Explainability, Privacy, Robustness
Secondary Category (if applicable):	Vendor Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation
Applied ML Model:	DOHA

When external users utilize a machine learning system, what those users are told about the decisions the system makes are of great importance. Indeed, there is no one right answer for the correct amount of information, and the act of choosing which information to reveal is a difficult balancing act. On the one hand, there is the need to

provide external users with accountability and transparency from public agencies. On the other hand, several competing factors may suggest less information be revealed:

- Issues of trade secrecy from the vendor (see *Vendor Principles #1*)
- The necessity to protect the machine learning system from adversarial attacks (see *Optimization Principle #4*)

For the DOHA model, users should have the right to see how the DOHA model performed if they were denied a clearance – did the administrative judge approve them, only for the DOHA model to deny them and cause a further review? Or did the opposite occur? This information is important enough for applicants that it should be released to them once the decision is made.

Principle 5: The “nudging” of external users by machine learnings systems should be carefully assessed and reviewed

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Human Autonomy, SDPR)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Deploy, Monitor, and Reassess
Applied ML Model:	None

While governments have long sought (either intentionally or not) to use a combination of incentives and behavioral psychology to affect citizen behavior, these ideas were first brought together and understood under the colloquialism of “nudging”

in 2008 (Thaler and Sunstein 2008). While nudging can be applied to any public agency system (AI or otherwise), nudging in the realm of machine learning systems can be particularly complex to deal with. As (Shrum, et al. 2019, 21) puts it, “[w]ith AI systems, a group of individuals can be provided certain information as a result of being identified by an AI system and ‘nudged’ to behave in a certain way or to believe certain things while other individuals are either not ‘nudged’ or are ‘nudged’ in a different direction”. It is easy to take this a step further and imagine such systems being used for potentially discriminatory purposes by nudging only a minority group in particular ways (Sunstein 2015). Some follow-on questions that arise from this principle, including:

Question 1: Where is the line when nudging becomes active manipulation? Is there such a line?

Question 2: Are this machine learning system’s external users likely to be particularly vulnerable?

Question 3: Do those responsible for designing and implementing user prompts have a strong incentive towards users providing a certain answer?

Principle 6: Internal users should be properly trained to both handle a ML model’s output and to consider their own possible biases towards AI

Responsible Actor(s):	Analyst
Relevant Research Thread(s):	Explainability, Democratic Legitimacy (non-algorithmic Fairness, Accountability, SDPR, Interpretability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Deploy, Monitor, and Reassess
Applied ML Model:	COMPAS

Beyond training to use the ML model's output itself, internal users responsible for making a public agency's assessments based on the output of a ML model should also be trained to understand two key issues they may be vulnerable to: decision-automation bias and automation-distrust bias. First, decision-automation bias is when internal users are "...hampered in their critical judgment, rational agency, and situational awareness as a result of their faith in the perceived objectivity, neutrality, certainty, or superiority of the AI system" (Leslie 2019, 21-22). In other words, it stems from an over-reliance on "the AI told me to do it", as though an AI's output should be automatically free of bias and immune from criticism or second thoughts. This idea is somewhat related to the famous Milgram shock experiment, which found that people were more likely to obey authoritative individuals in white lab coats into doing "bad" actions without applying critical thought (McLeod 2017).

Second, automation-distrust bias is at the opposite end of the spectrum, covering when internal users are inherently distrustful of an AI's output. The user will "disregard[s] its [the machine learning system's] salient contributions to evidence-based reasoning either as a result of their distrust or skepticism about AI technologies in general or as a result of their over-prioritisation of the importance of prudence, common sense, and human expertise" (Leslie 2019, 21-22). This is also inherently problematic – the ML model's outputs are meaningless if they're ignored entirely. As always, finding the proper balance is essential.

In the case of COMPAS, it is highly likely that the judges who rely (in part) on its assessments do not receive any special training on how to handle COMPAS' output. It is unclear if they are provided significant guidance on the potential pitfalls of using such a recidivism risk statistic in their analyses, and it is unclear if they are aware of their own potential positive or negative biases towards AI outputs.

6.6.4 Optimization Principles

The optimization principles below delve more deeply into the interaction effects between the six research threads (fairness, explainability, robustness, privacy, democratic legitimacy, and accuracy). An interaction effect is defined as a situation where optimizing for one element of an ANN (such as fairness) is likely to directly impact another element (such as accuracy). While there were more possible principles that could have been included in this section based on the findings of Stage One, I chose to be highly conservative in what was included due to the at-times lack of comprehensive empirical research into which research threads are in tension or complementary. Given how early this empirical scholarship is in its development, it would be too easy to add false principles that are not borne out by more extensive analyses. Additionally, because of these limitations and unlike the other sections of principles, many of these principles are more positive than normative in nature.

Principle 1: Constrained optimization inherently trades some level of optimization in one research thread in exchange for simultaneously optimizing one or more other research threads

Responsible Actor(s):	Manager
------------------------------	---------

Relevant Research Thread(s):	Accuracy, Fairness, Explainability, Robustness, Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	None
Development Stage(s):	Modeling, Testing, and Validation
Applied ML Model:	DOHA

Constrained optimization is where artificial limits are put in place during the ANN's training process to avoid certain "training paths" that might further optimize a given research thread because it violates a different research thread. One of the most powerful papers on the topic was put out by (Corbett-Davies, et al. 2017), where the authors showed that accuracy and fairness can very easily be at odds with one another – to reduce racial disparities in a given decision-making system, additional constraints were placed when optimizing for maximum accuracy. However, they showed that by doing so the final predictive accuracy was lower than it would have been without those constraints.

The same problem exists with adversarial examples – to make an ANN more robust against adversarial examples, the training data is sometimes modified in some form. This allows for easy conflict between the two optimization tasks – if optimizing for fairness requires one general training path and optimizing for adversarial examples requires another general training path, which should be taken? Alternatively, if one or the other optimization is applied sequentially, the one that is applied second may be forced to work with the training path already set out.

For the DOHA model, from previous principles we have seen several areas where constrained optimization may need to be considered. First, if there is access to more information about protected group statuses, constrained optimization may be needed to ensure algorithmic fairness. Second, since false positives are likely to be significantly less acceptable than false negatives, constrained optimization may be necessary to minimize false positives. However, both could decrease predictive accuracy or interfere with one another other if applied.

Principle 2: Different research threads should have varying difficulties in assessing optimization itself

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Accuracy, Fairness, Explainability, Robustness, Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Problem Formulation
Applied ML Model:	None

Some optimization problems are easier than others to *assess* in terms of how optimized they are at a given point. Based on my research, I have ranked them as such in terms of difficulty of assessing:

1. Accuracy
2. Privacy, Robustness, and Fairness
3. Explainability
4. Democratic Legitimacy

Keep in mind that the rankings above are not related to the difficulty of choosing the *correct* standard. Rather, I am attempting to answer regardless of if a given standard is

correct or not, how difficult is/are those standards to assess? Accuracy is clearly the easiest to assess – optimizing accuracy is the purest of numerical calculations – how accurately does a given ANN make a prediction? Regardless of whether you use F1 Score or BLEU or markedness or informedness, the math is simple, straightforward, and purely quantitative. Your choice of which accuracy metric to use may be wrong, but the actual calculations are not difficult regardless of which choice you make.

Next, privacy, robustness, and (algorithmic) fairness are roughly equal in terms of how difficult they are to assess. On the one hand, how optimized they are can be clearly defined quantitatively, just as accuracy can. However, at the same time they all have links to more qualitative concepts as well. In other words, they can also be linked directly to the functions of a public agency rather than the functions of the algorithm itself; even when the optimization of the math is done perfectly, the public agency's usage and implementation of that output can sometimes play a substantial role in the true difficulty of assessing these research threads.

Explainability is next, and its problem stems from the lack of easy quantifiability. As the literature review above discussed, how much explainability is sufficient and how does explainability technique X compare to explainability technique Y? There is no simple number that can be used to assess these, no matter which standard of explainability is chosen. Even excluding conceptions of interpretability (which deal with the public agency), explainability is still in part an inherently qualitative concept.

Finally, democratic legitimacy is clearly the most difficult and subjective research thread to “optimize” and assess. Simply put, there is no mathematical algorithm yet devised that can assess democratic legitimacy. It is an entirely qualitative task and a potentially subjective one.

Principle 3: Public agency managers should recognize that all definitions of algorithmic fairness cannot be met simultaneously

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Fairness, Democratic Legitimacy (Transparency, Deliberation, non-algorithmic Fairness)
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	None
Development Stage(s):	Problem Formulation Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

As has been noted previously several times in the literature review, all competing definitions for algorithmic fairness cannot be achieved simultaneously (Wadsworth, Vera and Piech 2018). Therefore, a public agency will inevitably have to pick and choose a definition (or a subset of different but compatible and overlapping definitions) that that agency will use for determining fairness. This is true for both algorithmic and non-algorithmic standards of fairness. While all standards will inevitably be imperfect and some will not be happy with any given definition of fairness, making the process public and transparent, as well as looking for stakeholder feedback, should help to meet the requirements of non-algorithmic fairness. Some of the discrimination-

related questions provided by (Shrum, et al. 2019, 20-21) are particularly relevant when making these assessments:

Question 1: How can AI systems be tested before they are employed to ensure that they will not discriminate among individuals in ways that have traditionally been prohibited or to determine if they are discriminating among individuals in unanticipated ways?

This has previously been discussed for the DOHA model.

Question 2: What redress or grievance procedures should be available to individuals who believe they have been unfairly treated as a result of an AI system?

DOHA itself already has an appeals process built in when someone is denied a security clearance – there is an appeal (which is this dataset), and there is even an appeal of the appeal (which is not included). Based on this, it appears there is already a strong grievance procedure in place to handle applicants who may feel wronged by the DOHA model. However, this could only be confirmed with internal information about DOHA.

Principle 4: The relationship between democratic legitimacy and robustness/privacy should be determined on a case-by-case basis

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Deliberation, Interpretability, Transparency), Robustness, Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Problem Formulation
Applied ML Model:	DOHA

Unlike fairness, accuracy, or explainability, where an increase in any of those three should always increase democratic legitimacy, increasing robustness and/or privacy may increase or decrease democratic legitimacy. This comes into play most prominently when issues of transparency, interpretability, and deliberation are at stake. On the one hand, optimizing for differential privacy and robustness would suggest minimizing how much information is provided to the public about how a given ANN was trained or created. However, doing so could harm democratic legitimacy by minimizing deliberation and reducing interpretability and transparency by having the public agency be more opaque about its internal workings.

In the case of the DOHA model, the primary worry here is with robustness (at least thus far). No one has yet de-anonymized clearance case summaries, but in the age of big data that may become possible down the road. However, in terms of robustness, applicants may attempt to “game the system” if they discover that particular kinds of answers to questions may cause the clearance case summaries to be written differently and thus interpreted differently by the DOHA model.

While this may be more difficult because the applicant’s input is indirect (the applicant does not write the clearance case summary), that does not mitigate the potential problem entirely. Complete transparency of all model outputs would potentially make this gaming more likely. Thus, democratic legitimacy and robustness are potentially in tension in this case, albeit not as strongly as in other potential cases.

Principle 5: De-anonymization techniques exist even when individual data has been made theoretically private

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Transparency, SDPR), Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial
Development Stage(s):	Problem Formulation Modeling, Testing, and Validation Deploy, Monitor, and Reassess
Applied ML Model:	DOHA

While this principle is admittedly less normative than the others, it still is important enough to include regardless. De-anonymization techniques are those which involve taking anonymized data and then determining the identity of the anonymous individuals with the help of external, oftentimes public datasets (Dorschel 2019) (Lee, et al. 2017). What this means is that anytime personally identifiable information is being actively used when training a machine learning system, this poses a potential issue in case the data is reidentified later on.

This is particularly true if that dataset is to be anonymized and made public on purpose for the sake of transparency. While all public agencies should seek for some level of transparency with the general public, this potential issue of de-anonymization makes it that much more complicated to determine how much transparency should be provided. The balancing act, then, is between privacy and transparency. Some questions that arise from this tension include:

Question 1: How should you calculate reidentification risk? Which standard of privacy is enough in which use cases?

The reidentification risk of applicants is very small (thus far) from the available data in the DOHA model. However, if more protected information were provided to the public (i.e. on race, ethnicity, etc.), this might increase the reidentification risk.

Question 2: Are there some fields of data which should be automatically removed simply because they are too dangerous to have be reidentified?

Not in the case of the DOHA model. Personal names are already removed, as are social security numbers.

6.6.5 Vendor Principles

The vendor is the key middleman that often exists for advanced software solutions in public agencies. Public agencies rarely have the resources to have their own expert internal data science team that can create complex ML systems, particularly state-of-the-art ANNs. Because of this, public agencies will often have a contract with a private firm that manages the ML system's development and deployment on its behalf. However, the usage of a private vendor come with additional issues to contend with. Of note, since the DOHA model is hypothetical and not created by a vendor, it is not discussed in this section.

Principle 1: Vendor claims to extensive trade secrecy should be treated with automatic skepticism

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (all), Explainability
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

Development Stage(s):	Problem Formulation Vendor Negotiations
Applied ML Model:	COMPAS

Significant research has been conducted at the intersection of intellectual property rights, machine learning systems, and public agencies plus our judicial system (Ram 2017) (Wexler 2017) (Shrum, et al. 2019) (Reisman, et al. 2018) (Brauneis and Goodman 2018). On the one hand, trade secrecy has inherent value to any free market society. Indeed, the ability for private firms to profit from their innovations is an essential element of the free enterprise in the US. On the other, “...it is unlikely that these [trade secrets] extend to information such as the existence of the system, the purpose for which it was acquired, or the results of the agency’s internal impact assessment” (Reisman, et al. 2018, 14).

In the case of ML systems, such secrecy can have a particularly high cost. First and foremost, there can be no transparency, explainability, or interpretability without at least *some* members of the public being aware of how a ML system functions in the first place, and accountability is also harmed as a result. Open records laws in many states make exemptions for trade secrets, and public agencies (at the federal, state, and local level) have in the past used this reasoning to prevent even basic transparency requests to their ML systems (Brauneis and Goodman 2018, 153-154).

Of course, COMPAS is where much of the foundations for this principle originate from. The fact that Northpointe never released their COMPAS model to public

inspection (or even to independent inspection by a select private group to protect trade secrecy) ensures that there are countless questions that simply cannot be answered.

Principle 2: Capabilities for maintenance should be transferrable where feasible

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Democratic Legitimacy (Algorithmic Maintainability)
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Vendor Negotiations
Applied ML Model:	COMPAS

If the ML system is controlled by an external vendor, there may come a day when that vendor is no longer wanted or needed, such as when the contract is transferred to a new private firm. In such cases, the public agency may want to ensure that there are contractual provisions such that the ML system can be transferrable to another vendor. Otherwise a public agency will be left with either being permanently stuck with a single vendor for a critical system, starting from scratch with a new vendor and a new ML system, stop using the ML system entirely, or enter a legal dispute.

In the case of COMPAS, it does not appear that there is such transferability to another vendor (although this remains unknown). If the Wisconsin judicial system considers COMPAS to be particularly important, this means that they will be unable to easily switch vendors and maintain functionality unless they wish to stop using COMPAS entirely. While such transferability is not always feasible (for example due to proprietary

technology and trade secrecy), public agency managers should consider the potential long-term reliance on one particular private firm that this causes.

Principle 3: Public agencies should carefully vet which external entities should be provided what level of data as a part of the vendor contract

Responsible Actor(s):	Manager
Relevant Research Thread(s):	Privacy, Democratic Legitimacy (Accountability, Transparency, Deliberation)
Secondary Category (if applicable):	General Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Problem Formulation Vendor Negotiations
Applied ML Model:	COMPAS

Building on *Public Agency Managers Principle #5* and *Vendor Principle #1*, attempting to add stakeholders after contracting with the vendor can be difficult due to issues of trade secrecy and contract requirements. Because of this, the stakeholders should be defined in procurement contracts specifying what kinds of data those stakeholders should be granted throughout development and implementation. These external groups may include non-governmental “good governance” organizations, independent researchers, legislative oversight committees, and the public in general, among others.

Issues of privacy, accountability, and transparency all intersect within this principle. First, there will be privacy concerns regarding any data transmitted to an external entity, both in terms of the vendor’s worry about losing trade secrets and the

public agency’s worry about personal data being unnecessarily revealed. Second, there will be transparency/accountability concerns given that it is unlikely that every external entity that wants access will be given access, or will be given as much access as it wants to the model and its data. In this case, privacy and accountability/transparency conflict with one another, and finding the proper balance behind them may be difficult.

In terms of COMPAS, while the firm Northpointe did release large quantities of training data and COMPAS’ assessments based on that training data to the public, it is unclear to me if they were contractually obligated to do so or if this was entirely voluntary and external users simply “got lucky”. Regardless, Northpointe still has not allowed *any* external entity (to my knowledge) to view their model itself, which should be seen as highly problematic.

Principle 4: Internal expertise requirements for evaluating vendor systems should be recognized as significant by public agencies

Responsible Actor(s):	Manager Developer
Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency, Deliberation, SDPR)
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full
Development Stage(s):	Vendor Negotiations
Applied ML Model:	None

Simply because someone is a public agency manager or a procurement specialist does not make them inherently qualified to determine which vendor system should be

designed or purchased when dealing with ML systems. Even though these principles are designed to help with those determinations, the decision stills require significant technical subject matter expertise. Therefore, it's important to consider what experience and expertise is required when evaluating a potential machine learning system before it is designed or developed (Shrum, et al. 2019, 20-21).

6.6.6 Conclusion

In conclusion, I believe that the principles above will help to provide a foundation for public agencies seeking to implement ANNs and other ML systems. However, the principles above should not be viewed as a finished product – rather, they are a starting point. There are still other fields of study, such as behavioral psychology (Russo 2018), that have valuable insights to add to and further develop and refine this framework. Indeed, there is also a great need for additional empirical research to determine bilateral relationships more strongly.

In the case of COMPAS, I would argue that based on the principles in this framework it should not continue to be used without substantial revisions, and even with the revisions below it is arguable simply that such black box predictive models do not belong in our judicial system. At the least, there is room for improvement in multiple areas:

- Ensuring that their contract does not preclude at least some sharing of their proprietary model

- Having a genuine public discussion and comment period about the best standard for algorithmic fairness, rather than have Northpointe determine such a standard themselves
- Assess the COMPAS model against potential competitors to determine if other ML models provided superior outcomes
- Ensure that COMPAS' recidivism assessment is provided to judges visually in a way that allows the judge to understand more about where the statistic comes from
- Create a model factsheet for COMPAS to increase transparency as to how it was designed and at least name the particular algorithm or group of algorithms used to predict with it

As to the viability of the DOHA model, I would conclude based on my framework's principles that overall it is not possible to mitigate enough of the issues to apply it in the real world. Despite the fact that attempting to follow these principles and answer the questions shed significant light into areas that could be improved to the DOHA model as-is, even if those improvements were made, the broader question of should it be used in the first place is still a definitive *no* in my view.

There are several key problems with using the DOHA model that using this framework revealed. First, the very fact that the model is based on such extraordinarily subjective data as the text of a judge's clearance case summary makes the model potentially unreliable. Something as simple as someone's unique writing style or future

administrative changes to clearance case summary formatting could completely change the effectiveness of the model.

Second, while the overall accuracy of the model is not inherently suspicious, the model makes extremely confident individual assessments; it generally surpasses 98% confidence for each individual assessment of a clearance case summary, whether granted or denied. In terms of explainability, this could imply that the model is “cheating” in some fashion (i.e. finding correlated words or phrases that have nothing to do with an applicant’s clearance suitability to determine whether it should be granted or denied). Otherwise, one would expect there to be less than high-absolute confidence in many of the model’s assertions.

Third, there is a lack of information about protected groups (beyond gender) to have any confidence that the model is treating those groups fairly, regardless of the fairness standard used. While the DOHA model’s extremely high accuracy might implicitly assure us that those applicants are highly likely treated fairly, there still should be more confidence that the model is fair to other protected groups as well.

Finally, there is the potential risk of applicants attempting to game the system if they knew that the way in which they phrased their answers might impact how the system assessed their suitability. This goes back to issues of robustness – the model is only ~97.5% accurate against the kinds of non-maliciously modified data it has seen

before. If adversarial examples were provided to it (even indirectly through the “finding of facts” from a judge/investigator), this might severely decrease accuracy.

In summation, these risks in my view are simply too great – I would recommend that DOHA not implement any kind of ANN or advanced ML system to assist in its determinations of clearance suitability at this point.

7 Conclusions

This study utilized a qualitative, multimethod approach consisting of archival research, expert interviews, peer review, and comparative analysis. The methods were combined to iteratively improve the final analytical framework. The result was an analytical framework consisting of 5 categories and 30 distinct principles. It is intended to provide public agency managers and analysts with guidance to assist them in their conception, development, and implementation of ANNs and other ML systems within their public agencies. Rather than trying to answer to every conceivable question that might need to be asked, its purpose is to help find the right questions that need to be asked.

7.1 Final Thoughts

To repeat what I wrote earlier, we stand at the precipice of a new kind of government for a wide array of public services. Utilizing machine learning, particularly artificial neural networks, has the potential for significant improvements to these services while also opening the door to new problems. Only by understanding how we got here and asking for assistance from fields as wide ranging as behavioral psychology, public administration, law, and computer science will we be able to ensure that the

positives outweigh the negatives. This analytical framework should not be seen as the end of the road, but a beginning for future research.

7.2 A New Case Study: Clearview AI and Facial Recognition

A new case at the intersection of machine learning and public policy has recently arisen in the US, but unfortunately it was too recent to be included in the main content of this dissertation. Nevertheless, it is relevant and important enough that it at least deserves mention here. Indeed, this case may be an even bigger “poster child” for the necessity of such a framework of principles than the case of COMPAS.

In January 2020, news reports from the New York Times came to light of a start-up company called Clearview AI and their facial recognition system (Hill 2020). Through gathering over 3 billion photos of individuals from social media websites across the internet, they have since contracted with over 600 law enforcement agencies worldwide to provide their facial recognition services to law enforcement agents through simple smartphone apps. Those agents can upload photos they take of suspects on those apps to Clearview AI’s servers, and Clearview AI then sends the law enforcement officer any matching faces their algorithm found from their massive database.

We can already see several problematic areas related to my framework above. First, when the New York Times reporter attempted to contact Clearview AI, he was originally unable to reach them. He soon (voluntarily) asked a police officer using Clearview AI’s app to upload his photo to see what it would return with; it came back

with several matches, which was unsurprising. However, “[a]fter the company realized I [the reporter] was asking officers to run my photo through the app, my face was flagged by Clearview’s systems” and it did not return matches any longer.

Clearview AI claimed that this was a software bug, but that is a difficult claim to take at face value considering that they are using a machine learning system that (if we remember) is unlikely to have human-made rules in its matching. However one feels about the idea of police using *en masse* facial recognition based on collected social media data, it is hard to dispute that a private firm probably shouldn’t have what appears to be essentially oversight powers over law enforcement queries. This raises substantial questions of accountability and transparency, to say the least.

Next there is the fact that there is no external verification on any Clearview AI system – their database cannot be verified for accuracy externally, and again (just like with COMPAS) the particular machine learning algorithm they’re using is proprietary. However, given that they are doing facial recognition and that multiple police sources have reported high levels of predictive accuracy, it is highly likely that they are employing some version of a neural network, probably a convolutional neural network.

Then there is the fact that some law enforcement agencies are relying solely on overall predictive accuracy to assess Clearview AI’s system. According to the former Indiana State Police captain who used Clearview AI’s system, “[f]or us, the testing [of Clearview AI’s app] was whether it worked or not.” (Hill 2020). It is understandable why

this perspective might exist among law enforcement; indeed, Indiana State Police achieved a positive match on a criminal within just 20 minutes of first using the app. However, such a perspective unfortunately papers over a whole host of potential problems related to transparency, fairness, accountability, deliberation, and privacy.

The list goes on; I can also see immediate further issues with the lack of transparency and interpretability – Clearview AI’s legal representative wrote a legal memo, which police appear to be following, stating that “authorities don’t have to tell defendants that they were identified via Clearview, as long as it isn’t the sole basis for getting a warrant to arrest them.” So aside from the other issues, defendants may not even know if this system was used against them. While Clearview AI violated the Terms of Service for various social media websites to collect their massive trove of data (which their founder readily admits), since all of the facial information is publicly available it appears as though they are violating no laws as of yet.

In summation, even from this quick and cursory analysis, we can see that there are significant problems with using Clearview AI’s system. It may indeed be able to produce extremely high levels of accuracy in its matches, but as I hope I have shown throughout this dissertation, that is insufficient for a public agency. It is unfortunate that this information did not come to light 6 months earlier or I might have rewritten this dissertation to focus on it rather than COMPAS.

7.3 Impact on Public Agency Behavior

The intent of this framework is to shape the behavior of public agencies (particularly public agency managers) when it comes to their utilization of ANNs and ML systems more broadly. Most human beings, including the employees of a public agency, will automatically turn to accuracy to see how effective a given AI is – after all, what could be more important than how often an AI correct? As I hope this framework has shown, however, the question of accuracy is only the very first surface-level question that needs to be asked. Indeed, even accuracy is subjective – the section on accuracy in the literature review makes clear that the accuracy statistic on its own may not tell the whole story when compared to recall, precision, etc.

Beyond accuracy, this framework urges public agency managers in particular to consider other key research threads. Generic platitudes about fulfilling fairness, privacy, robustness, explainability, and democratic legitimacy are replete in existing literature, but the principles in this framework attempt to provide more actionable ideas that public agency managers can follow to help achieve them. Based not only on what lessons have been learned from the past but what the past and present tell us about the future, this framework should help both managers and analysts alike better implement ANNs and ML systems more broadly in such a way that the public can have more trust in these systems. Perhaps most importantly, the material in this dissertation is more accessible to those without a deep specialization in computer science.

7.4 Addressing Literature Gaps

While the array of literature regarding ANNs continues to increase at a staggering rate, there are nevertheless gaps that need to be addressed. As this dissertation should show, there is already extensive literature regarding algorithmic governance and ethical AI issues generally. Additionally, issues of fairness, robustness, explainability, and privacy with regard to ANNs are not hard to find. However, with the exception of (Leslie 2019), there has been little to no *comprehensive* literature discussing how public agencies should utilize ANNs besides this dissertation.

Indeed, given the wide subject matter that this dissertation covered, it should be no surprise that there is substantial follow-on research that needs to occur both within the field of public policy and outside of it. First, there is a need for more experimental research and empirical evidence related to the bilateral relationship of different research threads.

Second, there needs to be more behavioral psychology research into how humans interact with ANNs and advanced ML systems more generally. This dissertation briefly touched on behavioral psychology through discussions of “nudging” and user interfaces, but that only skims the surface of these issues.

Finally, there needs to be more interdisciplinary computer science research into all areas of ANNs, particularly between public policy scholars and computer scientists. While I was able to identify potential trends from early research, it is possible that more

thorough and concrete studies in the future will disprove them. Aside from a few pairs of threads, such as fairness and accuracy or robustness and privacy, most threads only have relatively thin empirical research as to their bilateral relationship.

7.5 Relevance to the Future

Machine learning in general and artificial neural networks in particular are unlikely to dissipate as a relevant issue for public agencies to have to deal with. Until now, the approach has mostly been haphazard – public agencies would follow standard protocols for software development. However, these protocols have clearly fallen short – as the review on democratic legitimacy should make clear, there is already a substantial enough history of machine learning in public agencies to show that those agencies are unlikely to be following too many of the principles in this framework. Indeed, whether they are following any of these principles is *itself* difficult to know because of the near-total lack of transparency in some cases. This makes studying current usage extremely difficult, and was one reason why I eschewed attempting to study many other examples of ML systems used by public agencies already (with the exception of COMPAS and my own personally devised neural network).

In conclusion, it is my sincere hope that this dissertation is a clarion call for future research at the intersection of public policy and computer science. Only by looking at both sides will we be able to further develop best practices for public agencies.

Appendix A-1: Original Framework, Pre-Interviews

From the first two Stages of my research methodology arose this series of principles. Some of them were directly extracted from either the first or second stage, others were implicit in several distinct concepts, and still others were a combination of several different principles from different texts. Each of the principles below is not designed to be the final answer, but rather to provoke the right questions to be asked – those questions will have different answers for different use cases as techniques evolve and change, but the principles behind them should remain more constant. Some of the principles explicitly ask follow-on questions.

The principles were split into three categories: general principles, human interaction principles, and optimization principles.

A-1.1 General Principles

These are the broadest principles in this framework. Rather than being pulled explicitly from one document or another, they were principles that showed themselves, implicitly or explicitly, from several sources. Although several of these principles may seem obvious to a computer scientist, they are meant for usage by public agency managers who are much less likely to have that background.

Principle 1: Democratic legitimacy is intrinsically tied to each of the other research threads

While each of the other research threads can be defined separately from one another (i.e. it's not difficult to determine where measuring robustness ends and measuring fairness begins), democratic legitimacy is unique. It is pervasive and if defined broadly enough could theoretically encompass all other research threads within its paradigm. Because of this, and because of the inherent importance of achieving democratic legitimacy for any public agency, it should be the starting point for implementing ANNs in public agencies.

Principle 2: ANNs predict correlation, not causation

The principle "correlation does not imply causation" is well-known and should be self-evident. Nevertheless, it has special meaning in the context of ANN development. ANNs do not predict causation; rather, their predictions are only based upon countless subtle correlations. ANN output should never be thought of as implying a direct causation between the input and the output. This is particularly important considering that the end-users of a given ANN, whether it be the public or the average analyst at a public agency, are unlikely to be well-versed in statistical analysis. (Faes, et al. 2019)

Principle 3: Define your key definitions early and review them often for refinement

The myriad of taxonomies presented in this study should show both the incredible breadth of techniques that exist and the amount of critical terms with subjective and debatable definitions. More specifically, how you define (and justify your definitions) for fairness, explainability, robustness, democratic legitimacy, and privacy are critically important. Even accuracy requires definitions, such as whether recall,

precision, and F-1 score will be utilized. Several follow-on questions arise from this principle:

Question 1: Which definition(s) of fairness will you be optimizing for and why?

Question 2: Is differential privacy enough when defining privacy generally?

Question 3: How much explainability is sufficient for your ANN and why?

Question 4: How problematic are adversarial examples for your ANN and what level of robustness certification is necessary?

Principle 4: The importance of each research thread can vary between different ANNs and different use cases

Save for democratic legitimacy at the top, permanently ranking the research threads in importance is a poor idea for a public agency because how important each one is can vary from case to case. For example, if there is no interface through which malicious input could be provided to an ANN, optimizing for robustness becomes less important. Likewise, if the data involved isn't personally identifiable information and is already available to the public, the value of optimizing for privacy is also reduced. That's not to say that they are entirely unimportant in those instances, but simply that considering other research threads should be of greater importance.

Principle 5: Attempt traditional ML techniques before moving to ANNs

This is a principle which should be considered at the very beginning of any sort of analysis – is an ANN the best kind of ML model? Even if ANNs have achieved broadly superior results to more classical ML techniques in the realm of raw accuracy, the potential loss in the remaining research threads may indicate that ANNs are not the best solution.

Principle 6: Public agencies are unlikely to have intrinsic legal problems implementing ANNs but should still tread cautiously

According to a legal analysis conducted by (Coglianese and Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era* 2017), ANNs (and machine learning in general) shouldn't face any intrinsic legal problems when being implemented within a public agency. Nevertheless, there should still be due caution due to the potential for misuse and poor policy that may arise from an improper implementation. Despite the lack of legal limitations, or perhaps even because of them, how public agencies choose to use ANNs becomes increasingly important.

Principle 7: Unrepresentative data can exist even in cases where an entire population is the sample

I use the term "unrepresentative data," in contrast to "sampling bias," because an entire group or population can sometimes *be* the dataset when it comes to training ANNs. Under the traditional sampling bias paradigm, no sampling means no sampling bias. However, even when using an entire group or population as the dataset, the data may *still* be unrepresentative of reality. This issue is discussed in greater detail within the definition of fairness provided above. If your data isn't representative of reality, insights gleaned from an ANN trained on that data and with 99.9% "accuracy" may be tainted and produce unfair results.

Principle 8: ANNs require algorithmic maintenance

Algorithmic maintainability (as applied to ANNs) is the idea that there should be specific procedures in place to handle how the ANN should be periodically analyzed and potentially updated to ensure it is maintaining the same level of performance as it was

when previously checked. However, updating for the sake of updating should be avoided. Unlike with traditional software packages, a “new” version should not be construed as implying the system is inherently superior. Rather, such updates need to be considered on a case-by-case basis.

Several questions arise from this principle:

Question 1: Should new data be inputted continuously into a neural network’s training as it becomes available?

Question 2: Should new data be ignored until it is proven that the neural network’s performance is degrading?

Question 3: Should the ANN be retrained every N years from scratch?

Question 4: Should data from at least Y years old be removed from future iterations?

Question 5: What if new data causes accuracy to decline or bias to rise or some combination of the two?

Question 6: What human being is responsible for the decisions stemming from the ANN?

A-1.2 Human Interaction Principles

There are two potential human “audiences” for an ANN implemented in a public agency: internal users and external users. Internal users include members of the public agency themselves who would use those predictions as a part of their job, and external users are those in the general public that interact with the system. Some ANNs will only have an internal audience, some will have only an external audience, and some will have both. Regardless, the principles below focus on how humans interact with ANNs, who is responsible for that interaction, and what the effects of that interaction are.

Principle 1: User interface design is a critical feature and not an afterthought

This principle is emphasized by (Pásztor 2018) most prominently. The basic idea of this framework is that a computer system is not simply its input and output, it's also its *interface*. All too often, government systems have terrible user interfaces, and that may cause issues in the development and usage of ANNs. For example, a poor user interface for internal users may allow flaws and/or biases to go unnoticed. Likewise, a poor user interface for external users may provide an incorrect sense of what the system determined and why it determined it.

Principle 2: Enable internal users to do their own testing

The design of almost any ANN should allow for internal users to do their own testing and analysis. For example, these users should be able to enter fake data to help them get a better understanding of what the system is capable of and to spot potential design flaws. Even if internal users aren't computer scientists, they're likely to be subject matter experts and may be able to detect problems that would be otherwise missed.

Beyond simple testing, complex analytical suites are also available such as Stanford's open-source neural network verification project. (SyncedReview 2019) Such analytic suites are essential to help internal users grasp if an ANN is behaving as it should be even before external experts or consultants are hired to review it.

Principle 3: Be cautious of empathy loss through relying overmuch on ANNs

Sometimes the pure quantitative facts don't provide a complete picture of a given situation. Whereas decisions made by human beings have at least the chance of

human empathy allowing for exceptions in extreme situations, ANNs don't inherently allow for such exceptions. The impact of this can be edge cases where a human observer would be very likely to decide against standard policy, whereas an ANN will not.

Principle 4: Provide a model fact for an ANN in production

A model fact sheet is a standardized, relatively non-technical outline of the capabilities and limitations of a given ANN. Depending on the use case, the precise content can vary greatly. (Brajer, et al. 2019) provide an example of a model fact sheet in the case of healthcare delivery. Regardless of the use case, such a fact sheet should allow non-expert internal users to both understand how a model works and to have model-to-model comparisons.

Principle 5: The use of ANNs may compel agency decision makers to engage in quantitative coding of value judgments that have typically been made qualitatively

Public agencies often have qualitative values encoded into the decisions that the agency is responsible for. These qualitative values aren't always clear-cut, and can have a significant amount of subjective value assessment. One possible problem when trying to implement an ANN for such a case is that the ANN cannot accept these purely qualitative and often intuitive value assessments as inputs – it requires those inputs to be quantified to some degree. This change from qualitative to quantitative values may be difficult for some agencies to manage, and should be carefully scrutinized during development.

A-1.3 Optimization Principles

The optimization principles below delve more deeply into the interaction effect between the six research threads (fairness, explainability, robustness, privacy, democratic legitimacy, and accuracy). An interaction effect is defined as a situation where optimizing for one element of an ANN (such as fairness) is likely to directly impact another element (such as accuracy).

Principle 1: Constrained optimization inherently trades some level of optimization in one research thread in exchange for simultaneously optimizing one or more other research threads

Constrained optimization is where artificial limits are put in place during the ANN's training process to avoid certain "training paths" that might further optimize a given research thread because it violates a different research thread. One of the most powerful papers on the topic was put out by (Corbett-Davies, et al. 2017), where the authors showed that accuracy and fairness can very easily be at odds with one another – to reduce racial disparities in a given decision-making system, additional constraints were placed when optimizing for maximum accuracy. However, they showed that by doing so the final predictive accuracy was lower than it would have been without those constraints.

The same problem exists with adversarial examples – to make a ANN more robust against adversarial examples, the training data is often modified in some form. This allows for easy conflict between the two optimizations – if optimizing for bias requires one general training path and optimizing for adversarial examples requires

another general training path, which should be taken? Alternatively, if one or the other optimization is applied sequentially, the one that is applied second will be forced to work with the training path already set out.

Principle 2: Look beyond bilateral relationships

While this paper looked principally at the bilateral relationships between different research threads, this is just the beginning. There may be relationships between three or more research threads that only become apparent when all three are being optimized simultaneously. When developing an ANN, these relationships should be considered as well.

Principle 3: Different research threads have varying difficulties in assessing optimization itself

Some optimization problems are easier than others to *assess* in terms of how optimized they are at a given point. Based on my research, I have ranked them as such in terms of difficulty of assessing:

1. Accuracy
2. Privacy, Robustness, and Fairness
3. Explainability
4. Democratic Legitimacy

Accuracy is the easiest to assess – optimizing accuracy is the purest of numerical calculations – how accurately does a given ANN make a prediction? Next, privacy (specifically differential privacy), robustness, and fairness are roughly equal in terms of how difficult they are to assess. On the one hand, how optimized they are can be clearly defined quantitatively, just as accuracy can. However, at the same time they all lack a

universal definition for what precisely *should* be optimized for. As discussed above, there are multiple competing and inconsistent standards of how we define fairness. For robustness, should only an absolute robustness certification be measured, and for privacy, perhaps federated learning or secure enclaves is a better kind of privacy to optimize for than differential privacy. While each of these definitions for privacy, robustness, and fairness can be assessed quantitatively, the difficulty in assessing them lies in choosing the correct definition to optimize for.

Explainability is next, and its problem stems from the lack of easy quantifiability. As the literature review above discussed, how explainable is sufficiently explainable and how does explainability technique X compare to explainability technique Y? There is no simple number that can be used to assess these, no matter which standard of explainability is chosen.

Finally, democratic legitimacy is the most difficult and subjective research thread to “optimize” and assess. There is no mathematical algorithm yet devised that can assess democratic legitimacy, and the concept of legitimacy itself is at times only in the eyes of the beholder.

Principle 4: All definitions of fairness cannot be met simultaneously

No matter how the concept of fairness and bias is defined, all competing definitions for fairness cannot be achieved simultaneously. (Wadsworth, Vera and Piech 2018) Therefore, a public agency will inevitably have to pick and choose a definition (or a subset of different but compatible and overlapping definitions) that that agency will

use for determining fairness. While this will inevitably be imperfect and some will not be happy with any definition, making the process public and transparent should help to meet the requirements of democratic legitimacy.

Principle 5: Privacy and Robustness should be optimized jointly

Among all five research threads (plus accuracy), no two are as closely related as privacy and robustness. While they seek different goals, each deals with preventing malicious external actors from improperly manipulating the ANN. Current research shows that at a minimum optimizing them won't put them into conflict, and they may even have a mutually reinforcing relationship. (Phan, Vu, et al. 2019) (Phan, Thai, et al. 2019) (Lecuyer, Atlidakis, et al. 2019)

Principle 6: The relationship between democratic legitimacy and robustness/privacy should be determined by a case-by-case analysis

As a corollary to the preceding principle, robustness and privacy may have a negative relationship with democratic legitimacy. This is unlike fairness, accuracy, or explainability, where an increase in any of those three should always increase democratic legitimacy. However, optimizing for privacy and/or robustness may paradoxically end up harming democratic legitimacy. This comes into play most prominently when issues of transparency and public deliberation are at stake. On the one hand, optimizing for differential privacy and robustness would suggest minimizing how much information is provided about how a given ANN was trained or created. However, doing so would harm democratic legitimacy by minimizing deliberation and reducing transparency (and thus potentially constitutional protections).

Principle 7: Agencies should develop meaningful external researcher review processes to discover, measure, or track impacts over time (AIA)

Public review of new policies is a frequent feature of many public agencies.

However, such reviews become increasingly important in the case of ANN development.

In order to achieve democratic legitimacy in particular, the public needs to be confident that the public agency isn't violating any constitutional protections. To do this, external researchers should be permitted to assess the ANN and its impact over time. At the same time, this transparency must be weighed against the harm to robustness and/or privacy that may come from this level of intervention. A balance between the two should be struck, although where that balance is will change between projects.

Appendix A-2: Second Draft Framework, Post-Interviews/Expert Review/Revisions

The second draft of my research methodology is the analytical framework below. It was written after the completion of Stage Four but before Stage Five in my research methodology. Its principles come from a wide range of sources: some were found through archival research and my literature review, others came from peer review, and still more through expert interviews. Some of the principles explicitly ask follow-on questions, whereas others allow the reader to determine their own follow-on questions as needed.

At the beginning of each principle, I provide a small table showing the relevant research thread(s), the principle's secondary category (if applicable), and whether the principle is as applicable to machine learning in general as it is to artificial neural networks. The principles were divided into five categories: (1) public agency manager principles, (2) general technical principles, (3) human interaction principles, (4) optimization principles, and (5) vendor principles. These categories were determined from a combination of natural division points that appeared when discovering different principles as well as through expert interviews. The categories themselves are more fully defined below.

Additionally, some further definitions are required:

Vendor: An external private firm which is responsible for the technical development of a machine learning system for a public agency.

Internal User: Public agency employees (usually more junior individuals in the agency's hierarchy) who are responsible for the day-to-day usage of the machine learning system within the agency.

External User: Members of the public who interact with and/or are assessed in some manner by the machine learning system.

A-2.1 Public Agency Manager Principles

Public agency manager principles are defined as principles that public agency managers should constantly keep in mind and apply over the entire course of the development and implementation process. They do not end after a given stage of development finishes (i.e. after a vendor is chosen). Rather, they are constant principles that should be considered and reconsidered.

Principle 1: Clarify your key definitions early, and review them often for refinement

Relevant Research Thread(s):	Democratic Legitimacy (Transparency), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

While this principle may be self-explanatory, it bears further analysis. The myriad of taxonomies presented in this study should show both the incredible breadth of techniques that exist and the amount of critical terms with subjective and debatable definitions. How you define (and how you justify your definitions) for the critical concepts comprised within each research thread will play a substantial role in both the actual implementation of machine learning systems as well as how the public perceives that implementation. Several follow-on questions arise from this principle:

Question 1: How much explainability is enough for your machine learning system and why?

Question 2: How problematic are adversarial examples for your machine learning system and what level of robustness certification is necessary?

Question 3: What type(s) of privacy are you implementing (differential privacy, federated learning, etc.) and why?

Question 4: How does your agency define fairness and why was that particular definition chosen against other definitions?

Principle 2: The importance of each research thread varies between different types of machine learning and different use cases

Relevant Research Thread(s):	All
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

Permanently ranking research threads in importance is a poor idea for a public agency due to the changing importance of each thread from case to case. For example, if there is no interface through which malicious input could be provided, optimizing for robustness becomes less important. Likewise, if the data involved isn't personally

identifiable information and is already available to the public, the value of optimizing for privacy is also reduced. That's not to say that they are entirely unimportant in those instances, but simply that considering other research threads should be of greater importance.

Principle 3: Public agency decisionmakers may be compelled to engage in quantitative coding of value judgments that were typically made qualitatively

Relevant Research Thread(s):	Democratic Legitimacy (Accountability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

Public agencies often have qualitative values encoded into the decisions that the agency is responsible for. These qualitative values aren't always clear-cut and can have a significant amount of subjective value assessment. One possible problem when trying to implement machine learning systems for such cases is that these systems cannot take these purely qualitative (and often intuitive) value assessments as inputs – it requires those inputs to be quantified to some degree. This change from qualitative to quantitative values may be difficult for some agencies to manage and should be scrutinized during development and implementation.

Principle 4: Define your stakeholders and invite them for collaboration at various stages

Relevant Research Thread(s):	Democratic Legitimacy (Deliberation, Transparency)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

Public review of new regulations or policies is a frequent feature of many public agencies, such as the Federal Communications Commission. (FCC 2019) However, such reviews become increasingly important when machine learning systems come into play. This is because even with public or expert input, there can be a significant lack of explainability in the results of a machine learning system.

A-2.2 General Sociotechnical Principles

These principles include broad principles generally focused on the more technical aspects of machine learning system development. However, while they are technical in nature, understanding these principles and how to answer the questions they bring up is not a question just of computer science competency but rather an assessment of the values of the public agency and what it requires.

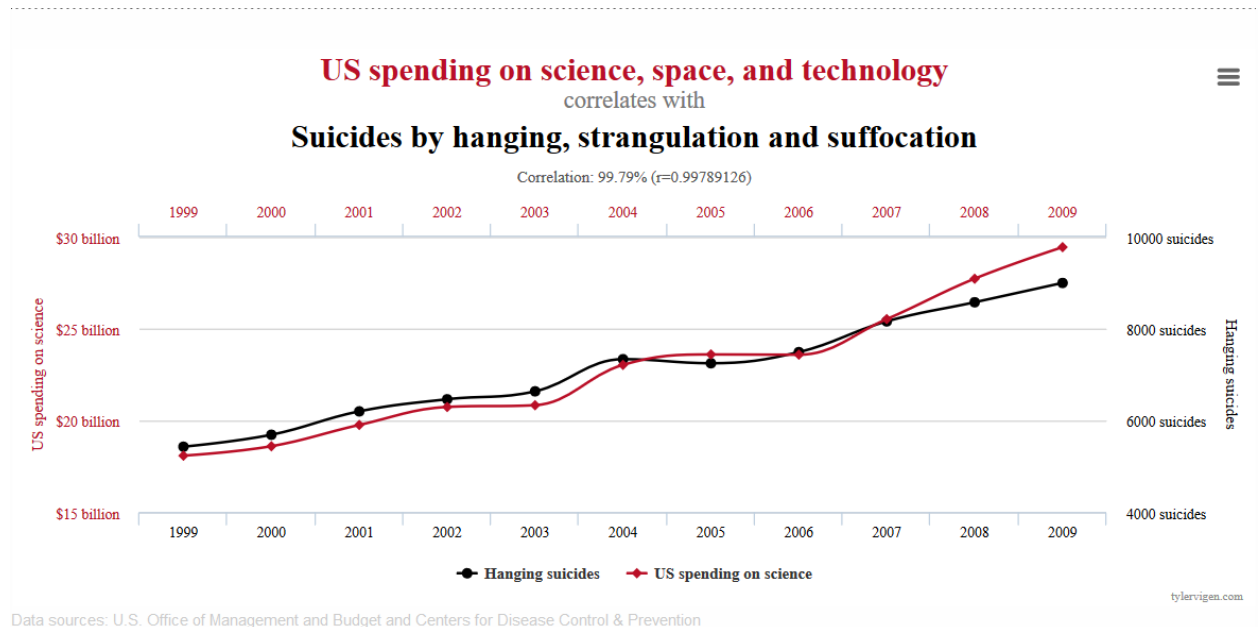
Principle 1: Correlation is not causation

Relevant Research Thread(s):	Fairness, Accuracy, Democratic Legitimacy (SDP)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

The statistical adage “correlation does not imply causation” is well-known and should be self-evident. (S. Singh 2018) Nevertheless, it has special meaning in the context of machine learning systems, particularly ANNs. Because of this, it deserves extra attention. Notably, machine learning systems do *not* predict causation; rather, their predictions are only based upon countless subtle correlations. ANN output should

thus never be thought of as determining a direct causation between the input and the output. This is particularly important considering that the end-users of a given ANN, whether it be the public or the average analyst at a public agency, may be unlikely to be well-versed in statistical analysis. (Faes, et al. 2019)

Examples abound on the gap between correlation and causation. Consider the following examples from Tyler Vigen's book/website *Spurious Correlations* (Vigen 2015):



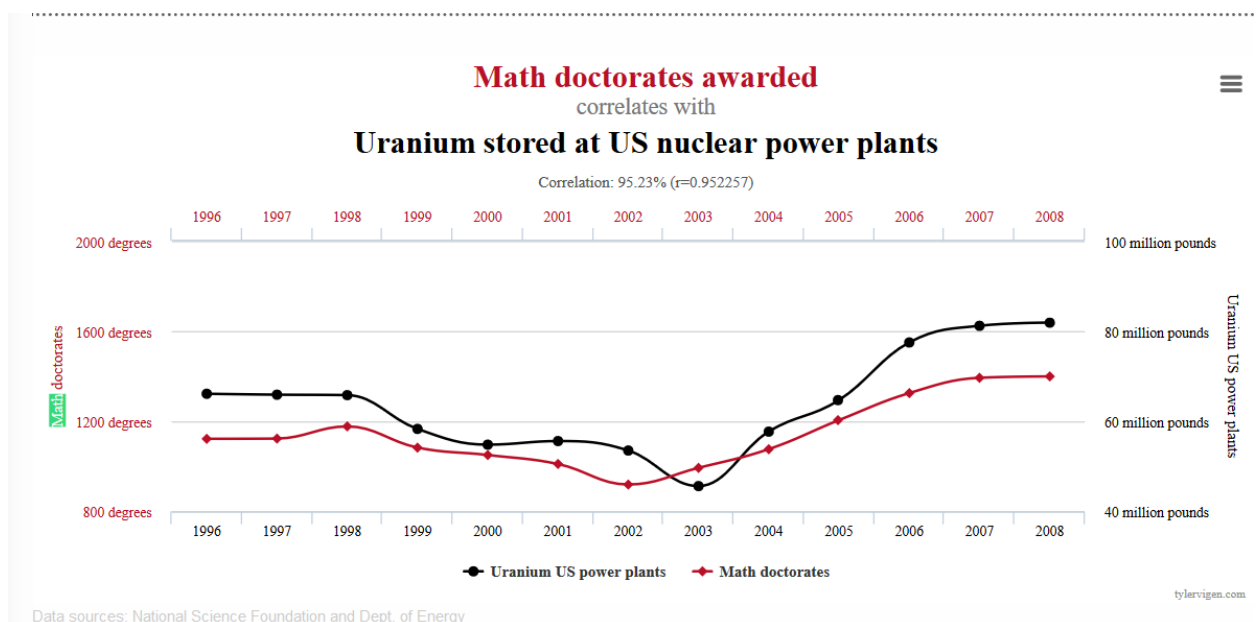


Figure 16 - Correlations $\neq$ Causations (Old Example)

It should be self-evident that there is clearly no *causal* relationship between math doctorates and uranium storage or U.S. science R&D spending and suicide rates, yet both of these relationships have a correlation of over 95%. What can this tell us about machine learning systems? Simple - if you provided a neural network with the math doctorates awarded each year as input and had it attempt to predict the uranium stored at US nuclear power plans as output, it would perform admirably well. However, the deeper question underlying this for public agencies is whether making determinations based upon a given correlation is a positive or a negative.

Question 1: Should a causal relationship between input and output be required (or at least asserted) prior to using a given variable as the input?

Question 2: What is the standard for determining causality in such cases?

Principle 2: Test multiple types of machine learning systems

Relevant Research Thread(s):	Fairness, Accuracy, Explainability, Robustness, Privacy
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Full

Even if ANNs have achieved broadly superior results to more classical ML techniques in the realm of raw accuracy, the potential loss in the remaining research threads may indicate that ANNs are not the best solution for every case. What's more, there are still some areas where other machine learning systems can still match their results in terms of raw accuracy. Thus, rather than settling on one particular type of system from the beginning, it's worthwhile to have multiple kinds of systems built. Different machine learning systems suffer from different flaws which may be more or less important in different use cases.

Principle 3: Unrepresentative data can exist even in cases where an entire population is the sample

Relevant Research Thread(s):	Fairness, Accuracy, Explainability, Robustness, Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

I use the term "unrepresentative data," in contrast to "sampling bias," because an entire group or population can *be* the dataset when it comes to training machine learning systems. Under the traditional sampling bias paradigm, no sampling means no

sampling bias. However, even when using an entire group or population as the dataset, the data may *still* be unrepresentative of reality. This issue is discussed in greater detail within the *Fairness* literature review section above. If your data isn't representative of reality, insights gleaned from a machine learning system trained on that data, even with 99.9% "accuracy", may be tainted and produce unfair results.

Principle 4: Algorithmic Maintenance needs are higher than with traditional software algorithms

Relevant Research Thread(s):	Democratic Legitimacy (Algorithmic Maintenance, Accountability)
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

Algorithmic maintainability (as applied to ANNs in particular, though somewhat applicable to machine learning systems generally) is the idea that there should be specific procedures in place to handle how these systems should be periodically analyzed and potentially updated to ensure that they maintaining the same level of performance as when they were initially assessed. Unlike traditional software packages, updating a machine learning system automatically just for the sake of updating it may not be the correct solution. This is because unlike traditional software packages, a "new" version should not be construed as implying the system is inherently superior. Rather, what that update would entail is of great importance. Rather, such updates need to be considered on a case-by-case basis.

Several questions arise from this principle:

Question 1: Should new data be inputted continuously into an ANN's training as it becomes available, or should it be done with "versions" similar to traditional software development methods?

Question 2: Should new data be ignored until it is proven that the ANN's performance is degrading?

Question 3: Should the ANN be retrained every N years from scratch?

Question 4: Should data from at least Y years old be removed from future iterations?

Question 5: What if new data causes accuracy to decline or bias to rise or some combination of the two?

Question 6: What human being is responsible for the decisions stemming from the ANN?

Principle 5: Measuring "accuracy" is more than just the raw accuracy statistic

Relevant Research Thread(s):	Accuracy
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Full

As the *Accuracy* section in the literature review should have made clear, even the study of accuracy itself can be subjective. Public agency managers and internal users both need to be aware of what kinds of "accuracy" numbers a machine learning system is providing. The difference between assessing actual raw accuracy and F-1 Score can be substantial when considering the suitability of a machine learning system for implementation.

Principle 6: Minority groups within training datasets are in greater danger of poor model performance

Relevant Research Thread(s):	Democratic Legitimacy (SDP), Accuracy, Fairness
Secondary Category (if applicable):	Optimization Principles
Non-ANN Machine Learning Applicability:	Partial

Minority groups can be intrinsically at greater risk of unfair treatment by a public agency because there often won't be as much training data for the minority group. The correlation between the quantity of high-quality training data and model performance is extremely strong; all other things being equal, a model with significantly more training data will almost always achieve equal or (more often) superior performance to an equivalent model with less training data. Because of this, minority groups may be particularly vulnerable to a dearth of training data. While this may also be the case for other kinds of machine learning, it can vary between types.

Question 1: Does your training dataset have relevant minority sub-groups? If so, are you testing the predictive accuracy for those groups separately?

Question 2: Are there likely to be new minority groups with specific unique characteristics that are presently missing from the training dataset?

Principle 7: Question what you think is being learned

Relevant Research Thread(s):	Accuracy, Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

Machine learning systems, and particularly ANNs, will constantly try to “cheat” during training. By cheat, I meant that they will by design seek the easiest training path to maximize accuracy. For image recognition tasks, this often (though not always) means that the most obvious and consistent differences between two different image

classifications will be what the system learns. While this may seem like a good thing at first glance (and it often is), it isn't always a positive.

Indeed, there are several famous cases in the history of machine learning where this misunderstanding of what was learned caused substantial real-world problems. Perhaps the most well-known case is when a neural network was trained to differentiate between wolves and huskies. (Ribeiro, Singh and Guestrin 2016) On the one hand, the model achieved a very high accuracy during training and it was thought of as a great success initially (particularly back in 2016).

However, when it was applied to the real world, it failed spectacularly, misclassifying what should have been easy identifications between wolves and huskies. When researchers dug into the ANN, they found out that they had been wrong about what it had succeeded in classifying: the ANN hadn't been classifying the animals, but rather it had become adept at identifying images with snow in them. Since all the images with wolves had snow in them, it was thus able to successfully classify the wolves rather easily. In the real world, however, the images of wolves didn't always have snow in them. This caused the classifier to essentially malfunction. (Kepler 2019)

A-2.3 Human Interaction Principles

There are two potential human "audiences" for an ANN implemented in a public agency: *internal users* and *external users*. These terms were defined at the beginning of this section. Some ML systems will only have internal users and some will have both

internal and external users; it's extremely unlikely that an ML system will have no internal users at all. Internal or external, the principles below focus on how humans interact with ML systems.

Principle 1: User interface and user experience (UI/UX) are critical features, not afterthoughts, to machine learning systems

Relevant Research Thread(s):	Democratic Legitimacy (Accountability), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

This principle is emphasized by (Pásztor 2018) most prominently: a computer system is not simply its input and output, it's also its *interface*. All too often, government systems have truly *terrible* user interfaces (Sinders 2018), and that can cause significant issues in the development and usage of ANNs in particular. These issues can be for internal users and external users both. For example, a poor user interface for internal users may allow flaws and/or biases to go unnoticed. Likewise, a poor user interface for external users may provide an incorrect sense of what the system determined and why it determined it. While government has traditionally had a poor history of UI/UX, the problems arising from ANNs will not just be annoying, but can have significant negative real-world consequences.

Principle 2: Enable internal users to do their own testing

Relevant Research Thread(s):	Democratic Legitimacy (Maintenance), Explainability
Secondary Category (if applicable):	General Technical Principles

Non-ANN Machine Learning Applicability:	Full
--	------

The design of almost any ANN should allow for internal users to do their own testing and analysis. For example, these users should be able to enter fake data to help them get a better understanding of what the system is capable of and to spot potential design flaws. Even if internal users aren't computer scientists, they're likely to be subject matter experts and may be able to detect problems that would be otherwise missed.

Beyond simple testing, complex analytical suites are also available such as Stanford's open-source neural network verification project. (SyncedReview 2019) Such analytic suites are essential to help internal users grasp if an ANN is behaving as it should be even before external experts or consultants are hired to review it.

Principle 3: Create a model fact sheet for all internal users

Relevant Research Thread(s):	Democratic Legitimacy (Transparency, Accountability), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

A model fact sheet is a standardized, relatively non-technical outline of the capabilities and limitations of a given ANN. Depending on the use case, the precise content can vary greatly. (Brajer, et al. 2019) provide an example of a model fact sheet in the case of healthcare delivery. Regardless of the use case, such a fact sheet should

allow non-expert internal users to both understand how a model works and to have model-to-model comparisons.

Principle 4: Internal user acceptance of False Positives and False Negatives are not always equivalent

Relevant Research Thread(s):	Democratic Legitimacy(Accountability), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

Depending on the use case, internal users may be more or less receptive to false positives versus false negatives. In other words, there can be a significant difference between stating that a wrong answer is right versus stating that a right answer is wrong. This is entirely dependent upon the use case, however. For example, consider the case of an image classification system designed to track poachers. For those internal users assigned to understand what the system is saying, there is a significant difference in user acceptability between showing too many false positives (that is, showing that poaching was occurring when it actually wasn't) and false negatives (showing that poaching was not occurring when it actually was). Even though the raw accuracy may be the same regardless of whether the wrong answers are false positives or false negatives, the internal users in this hypothetical case are much more likely to be willing to get false positives rather than false negatives: better to sift through the false positives to find the real cases of poaching rather than miss actual cases of poaching entirely (within reason).

Principle 5: Determine what information should and should not be provided for external users

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency), Explainability, Privacy, Robustness
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

When external users utilize a machine learning system, what those users are told about the decisions the system makes are of great importance. Indeed, there is no one right answer for the correct amount of information. The act of choosing which information to reveal is a difficult balancing act. On the one hand, there is the need to provide external users with accountability and transparency from public agencies. On the other hand, several competing factors may suggest less information be revealed:

- Issues of trade secrecy from the vendor (see *Vendor Principles #1*)
- The necessity to protect the machine learning system from adversarial attacks (see *Optimization Principle #4*)

Principle 6: Determine if special redress procedures are needed for external users

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Deliberation), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

As with many principles here, there is no singular answer as to whether external users should be granted what I call *special redress procedures*, or in other words, a

particular procedure that external users can follow when they believe they have been wrongly treated by the machine learning system's determination. Current research on the subject varies in its conclusions between this simply being a necessary internal question for a public agency to consider (Shrum, et al. 2019, 20-21) to this being a mandatory prerequisite for any machine learning system in a public agency. (Reisman, et al. 2018) Some of the following questions may help to determine if such procedures are necessary:

Question 1: What is the overall importance of the determination? Does it have substantial reputational or financial implications?

Question 2: Does the system have significant PII as its input?

Question 3: How likely is it that the machine learning system's data that it has on an individual is flawed in some way?

Principle 7: The "nudging" of external users by machine learnings systems should be carefully assessed

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Autonomy), Explainability
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Full

The concept of "nudging" is usually thought to have originated from (Sunstein 2015). While it can be applied to any public administration system (AI or otherwise), nudging in the realm of machine learning systems can be particularly complex to deal with. As (Shrum, et al. 2019, 21) puts it, "[w]ith AI systems, a group of individuals can be provided certain information as a result of being identified by an AI system and

“nudged” to behave in a certain way or to believe certain things while other individuals are either not “nudged” or are “nudged” in a different direction”. It is easy to take this a step further and imagine such systems being used for potentially discriminatory purposes by nudging only a minority group in particular ways. Some follow-on questions that arise from this principle include:

Question 1: Where is the line when nudging-based manipulation occurs? Is there such a line?

Question 2: Are this machine learning system’s external users likely to be particularly vulnerable?

Question 3: Do those responsible for designing the user prompt have a strong incentive towards users providing a certain answer?

A-2.4 Optimization Principles

The optimization principles below delve more deeply into the interaction effect between the six research threads (fairness, explainability, robustness, privacy, democratic legitimacy, and accuracy). An interaction effect is defined as a situation where optimizing for one element of an ANN (such as fairness) is likely to directly impact another element (such as accuracy).

Principle 1: Constrained optimization inherently trades some level of optimization in one research thread in exchange for simultaneously optimizing one or more other research threads

Relevant Research Thread(s):	Accuracy, Fairness, Explainability, Robustness, Privacy
Secondary Category (if applicable):	General Technical Principles
Non-ANN Machine Learning Applicability:	None

Constrained optimization is where artificial limits are put in place during the ANN’s training process to avoid certain “training paths” that might further optimize a given research thread because it violates a different research thread. One of the most powerful papers on the topic was put out by (Corbett-Davies, et al. 2017), where the authors showed that accuracy and fairness can very easily be at odds with one another – to reduce racial disparities in a given decision-making system, additional constraints were placed when optimizing for maximum accuracy. However, they showed that by doing so (and meeting two of their standards for accuracy) the final predictive accuracy was lower than it would have been without those constraints.

The same problem exists with adversarial examples – to make a ANN more robust against adversarial examples, the training data is often modified in some form. This allows for easy conflict between the two optimizations – if optimizing for bias requires one general training path and optimizing for adversarial examples requires another general training path, which should be taken? Alternatively, if one or the other optimization is applied sequentially, the one that is applied second will be forced to work with the training path already set out.

Principle 2: Different research threads have varying difficulties in assessing optimization itself

Relevant Research Thread(s):	Accuracy, Fairness, Explainability, Robustness, Privacy
Secondary Category (if applicable):	General Technical Principles
Non-ANN Machine Learning Applicability:	Partial

Some optimization problems are easier than others to *assess* in terms of how optimized they are at a given point. Based on my research, I have ranked them as such in terms of difficulty of assessing:

5. Accuracy
6. Privacy, Robustness, and Fairness
7. Explainability
8. Democratic Legitimacy

Keep in mind that the rankings above are not related to the difficulty of choosing the *correct* standard. Rather, regardless of if a given standard is correct or not, how difficult is/are those standards to assess? Accuracy is clearly the easiest to assess – optimizing accuracy is the purest of numerical calculations – how accurately does a given ANN make a prediction? Regardless of whether you use F1 Score or not, the math is simple, straightforward, and purely quantitative. Next, privacy, robustness, and fairness are roughly equal in terms of how difficult they are to assess. On the one hand, how optimized they are can be clearly defined quantitatively, just as accuracy can. However, at the same time they all lack a universal definition for what precisely *should* be optimized for. As discussed above, there are multiple competing and inconsistent standards of how we define fairness. For robustness, should only an absolute robustness certification be measured, and for privacy, perhaps federated learning or secure enclaves is a better kind of privacy to optimize for than differential privacy. While each of these definitions for privacy, robustness, and fairness can be assessed

quantitatively, the difficulty in assessing them lies in choosing the correct definition to optimize for.

Explainability is next, and its problem stems from the lack of easy quantifiability. As the literature review above discussed, how explainable is sufficiently explainable and how does explainability technique X compare to explainability technique Y? There is no simple number that can be used to assess these, no matter which standard of explainability is chosen.

Finally, democratic legitimacy is the most difficult and subjective research thread to “optimize” and assess. There is no mathematical algorithm yet devised that can assess democratic legitimacy, and the concept of legitimacy itself is at times only in the eyes of the beholder.

Principle 3: All definitions of fairness cannot be met simultaneously

Relevant Research Thread(s):	Fairness
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	None

No matter how the concept of fairness and bias is defined, all competing definitions for fairness cannot be achieved simultaneously. (Wadsworth, Vera and Piech 2018) Therefore, a public agency will inevitably have to pick and choose a definition (or a subset of different but compatible and overlapping definitions) that that agency will use for determining fairness. While this will inevitably be imperfect and some will not be

happy with any definition, making the process public and transparent should help to meet the requirements of democratic legitimacy. Some of the discrimination-related questions provided by (Shrum, et al. 2019, 20-21) are particularly relevant when making these assessments:

Question 1: How can AI systems be tested before they are employed to ensure that they will not discriminate among individuals in ways that have traditionally been prohibited or to determine if they are discriminating among individuals in unanticipated ways?

Question 2: What redress or grievance procedures should be available to individuals who believe they have been unfairly treated as a result of an AI system?

Principle 4: The relationship between democratic legitimacy and robustness/privacy should be determined by a case-by-case analysis

Relevant Research Thread(s):	Democratic Legitimacy (all), Robustness, Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

As a corollary to the preceding principle, robustness and privacy may have a negative relationship with democratic legitimacy. This is unlike fairness, accuracy, or explainability, where an increase in any of those three should always increase democratic legitimacy. However, optimizing for privacy and/or robustness may paradoxically end up harming democratic legitimacy. This comes into play most prominently when issues of transparency and public deliberation are at stake. On the one hand, optimizing for differential privacy and robustness would suggest minimizing how much information is provided about how a given ANN was trained or created.

However, doing so would harm democratic legitimacy by minimizing deliberation and reducing transparency (and thus potentially constitutional protections).

Principle 5: De-anonymization techniques exist even when individual data has been made theoretically private

Relevant Research Thread(s):	Democratic Legitimacy (Transparency, SDP), Privacy
Secondary Category (if applicable):	N/A
Non-ANN Machine Learning Applicability:	Partial

De-anonymization techniques are those which involve taking anonymized data and then determining the identity of the anonymous individuals with the help of external, oftentimes public datasets. What this means is that anytime personally identifiable information is being actively used when training a machine learning system, this poses a potential issue in case the data is reidentified later on. While all public agencies should seek for some level of transparency with the general public, this potential issue of anonymized data being reidentified makes it that much more complicated to determine how much transparency should be provided. The balancing act, then, between privacy and transparency becomes the essential issue. Some questions that arise from this tension include:

Question 1: How should we calculate reidentification risk? Which standard of privacy is sufficient in which use cases?

Question 2: Are there some fields of data which should be automatically removed simply because they are too dangerous to have be reidentified?

A-2.5 Vendor Principles

The vendor is the key middleman that often exists for advanced software solutions in public agencies. Public agencies rarely have the resources to have their own expert internal data science team that can create these systems. Because of that, public agencies will often have a contract with a private firm that manages the software's development and deployment. However, with usage of a private vendor come with new issues that must be deal with

Principle 1: Vendor claims to extensive trade secrecy should be treated with caution

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency, Maintainability, SDP), Explainability, Privacy
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

There is significant evidence that one of the biggest tensions during ANN development is the tension between a vendor's trade secrecy and the relevant research threads noted above. On the one hand, trade secrecy has inherent value to any free market society. However, in the case of ANNs that secrecy has a particularly high cost. There can be no transparency (and thus no explainability or interpretability) without at least some members of the public being aware of how the ANN functions in the first place, which is prevented with trade secrecy. What's more, accountability is intrinsically limited as well. Even individual privacy can be harmed since it can be more difficult to categorically confirm how the public's data is being used.

Principle 2: Capabilities for maintenance should be transferrable

Relevant Research Thread(s):	Democratic Legitimacy (Maintenance)
Secondary Category (if applicable):	Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

This principle derives directly from maintainability – if the ANN is controlled by an external vendor, there may come a day when that vendor is no longer wanted or needed. The public agency may develop the capabilities to internally manage the ANN, a vendor may go out of business, or a new vendor which has lower prices may be sought after some period of time. Regardless of the reason, the ANN should be able to be transferrable from one firm to another, lest a public agency become permanently stuck and reliant upon a single vendor for a critical system.

Principle 3: Data provenance should be maintained as metadata

Relevant Research Thread(s):	Democratic Legitimacy (Transparency, Accountability), Explainability
Secondary Category (if applicable):	General Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

ANNs can have data arrive from a wide array of sources. That data can be split, rearranged, transformed, merged, reconstructed, and extrapolated from multiple times from initial data ingest to training. The determination of where this data came from, whether it's accurate, and how it's used are questions of significant importance when public agencies implement machine learning systems. (Shrum, et al. 2019, 19) Indeed,

maintaining specific and precise logs for where this data comes from and how it has been transformed since then becomes extremely important. Interpretation of the output of an ANN requires to some degree the ability to determine where the data came from and how valid that data is. Likewise, external reviewers require data provenance to ensure public agency accountability.

Principle 4: Determine which external entities should be provided what level of data as a part of the vendor contract

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency)
Secondary Category (if applicable):	General Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

Building on *Public Agency Managers Principle #4* and *Vendor Principle #1*, attempting to add stakeholders after contracting with the vendor can be difficult due to issues of trade secrecy and contract requirements. Because of this, the stakeholders should be defined early and specifying what kinds of data those stakeholders should be granted throughout development and implementation. These entities may include external model auditing groups, non-governmental “good governance” organizations, and the public in general, among others.

Principle 5: Evaluation of vendor systems can require extensive internal expertise

Relevant Research Thread(s):	Democratic Legitimacy (Accountability, Transparency, Deliberation, SDP)
Secondary Category (if applicable):	General Public Agency Manager Principles
Non-ANN Machine Learning Applicability:	Full

Simply because someone is a public agency manager or a procurement specialist does not make them inherently qualified to determine which vendor system should be purchased. Even though these principles are designed to help with those determinations, the decision may require significant subject matter expertise beyond these principles. Therefore, it's important to consider what experiences and expertise are required when evaluating a machine learning system before it is used. (Shrum, et al. 2019, 20-21)

Appendix B: An Example Artificial Neural Network in Public Policy

The purpose of this section is two-fold. First, it will walk the reader through the general process of creating an ANN at an abstract level. Second, it will provide a valuable, usable ANN to help understand and apply the principles noted in *Section 6.6*. The model produced here is referred to frequently in that section as the “DOHA model”. For a deeper inspection of the code used to produce the model, you can view the associated GitHub repository.¹⁰ Within this section, I will loosely follow the development stages outlined in *Section 6.5.16*, although all cannot be realistically achieved in a hypothetical case.

1. Problem Formulation
2. Vendor Negotiations
3. Data Extraction & Acquisition
4. Data Pre-Processing
5. Modeling, Testing, and Validation
6. Deploy, Monitor, and Reassess

B.1 Problem Formulation

Since this is a hypothetical example, I could draw upon almost any relevant data source and situation. The source I chose is generally well-known to those working in public policy: that of security clearances. I wanted to answer a simple question: is it

¹⁰ <https://github.com/Starstorm/Dissertation>

possible to develop an ANN that could predict whether someone should be granted a security clearance based on the textual content of their *clearance case summary*?

People working for the federal government often need security clearances for a wide variety of jobs in federal agencies. Among them, the largest federal agency with cleared employees is the Department of Defense (DoD). When DoD contractors have their security clearance applications rejected initially, they can appeal that decision through the Defense Office of Hearings and Appeals (DOHA).

The potential uses for this model are straightforward. Rather than attempting to *replace* administrative judges in DOHA, this model could be used for quality assurance purposes: if the model's accuracy is high enough, in theory it could be used as another "check" alongside an administrative judge's decision. For example, if the administrative judge says "clearance should be granted" but the ANN says "clearance should not be granted", this might theoretically cause an additional layer of review for the case.

Critically, this should not be seen as an endorsement of using such a model for this situation. Indeed, that is the purpose of the analytical framework in *Section 6.6*: to help determine not only how to best implement such a model, but also to answer the more fundamental question of whether or not it's even appropriate for an ML model to be built in the first place or whether the problems with developing and implementing the model are simply too great to mitigate.

B.2 Vendor Negotiations

Since this is a hypothetical example, there is no vendor to negotiate with.

B.3 Data Extraction and Acquisition

When an individual appeals their initial acceptance or rejection of a security clearance, DOHA kindly places all of their (anonymized) clearance case summaries going back to 1998 online, albeit not in spreadsheet format; web scraping was required to obtain the cases from their website and organize them in a table. The ANN was initially trained from about 5,000 of these clearance case summaries. This does not include all the clearance case summaries from the DOHA website, however – some of them were unable to be properly scraped, others did not have a decision clearly marked, and still more were actually second-layer appeals, which were excluded. The ground truth for this dataset will be the actual decisions by DOHA’s administrative judges on whether the security clearance should be granted or not.

B.3.1 Example Input Data

Below is an example of one of the input’s (from the roughly 5,000 in the dataset). It is an anonymized clearance case summary. Only the *Statement of the Case* and *Findings of Fact* were included. Note that formatting (such as spaces and newline characters) may be distorted from the original text:

96-0522.h1

December 31, 1996

In Re:

SSN:

Applicant for Security Clearance

ISCR OSD Case No. 96-0522

DECISION OF ADMINISTRATIVE JUDGE

MICHAEL KIRKPATRICK

Appearances

FOR THE GOVERNMENT

Earl C Hill, Jr., Esq.

Department Counsel

FOR THE APPLICANT

Pro Se

STATEMENT OF THE CASE

On July 30, 1996, the Defense Office of Hearings and Appeals (DOHA), pursuant to Executive Order 10865 and Department of Defense Directive 5220.6

(Directive), dated January 2, 1992, issued the attached Statement of Reasons (SOR) to (Applicant), which detailed reasons why DOHA

could not make the preliminary affirmative finding under the Directive that it is clearly consistent with the national interest to grant or continue a security clearance for the Applicant, and which recommended referral to an Administrative Judge to determine whether clearance should be denied or revoked.

Applicant responded to the SOR in writing on August 3, 1996, and in his Answer he elected to have the case determined on a written record in lieu of a hearing.

Department Counsel submitted the Government's File of Relevant Material (FORM) to Applicant on September 25, 1996. The Government submitted seven

items in support of its contentions. Applicant was instructed to submit information in rebuttal, extenuation or mitigation within 30 days of receipt. The date on

which Applicant received the FORM cannot be determined from the file, but he submitted additional material for consideration on October 25, 1996. On

November 20, 1996, Department Counsel submitted his written objections to the Applicant's additional material. Nevertheless, the undersigned Administrative

Judge has overruled Department Counsel's objections and considered Applicant's additional material submitted on October 25, 1996.

The case was assigned to the undersigned Administrative Judge on November 25, 1996.

FINDINGS OF FACT.

In his Answer to the SOR, Applicant admitted the material facts alleged in SOR subparagraphs 1.a., 1.b., 1.c., and 1.d., and those admissions are hereby

incorporated herein as findings of fact. The following additional findings of fact are entered as to each paragraph and subparagraph in the SOR:

Applicant is 34 years old, and he is employed as a ----- by a defense contractor.

A secret-level Department of Defense security clearance is

required in order for him to perform his assigned duties.

Paragraph 1 (Criterion H - Drug Involvement). The Government alleges that Applicant is ineligible for clearance because he has used marijuana and cocaine.

Applicant first smoked marijuana in 1978, when he was in high school. From 1980 to 1984, when he was in college, Applicant smoked marijuana from two to three times per month, on the average, although there was periods of up to four or five months when he abstained from smoking marijuana. (Items 3, 4, and 5.)

During this period of time, Applicant purchased marijuana once or twice, paying less than \$20.00 for an eighth of an ounce of marijuana on each occasion.

(Items 3, 4, and 5.) His motivation for smoking marijuana was "enjoyment and recreation." (Item 5.)

In March of 1983, on Applicant's 21st birthday, he snorted cocaine. He does not intend to use cocaine again. (Items 3, 4, and 5.)

On June 13, 1985, Applicant signed and submitted a Personnel Security Questionnaire (PSQ) as part of an employment and security clearance application process, certifying that his answers were true and complete and accurate. In that PSQ, Applicant stated that his desire and his opportunity to smoke marijuana had "dropped, though not disappeared." (Item 4.)

In his signed, sworn statement dated January 2, 1986, Applicant stated, "I have no intention of any future use of marijuana as it is not part of my current lifestyle." (Item 5.)

Nevertheless, Applicant did smoke marijuana during the period from December of 1989 to January 15, 1996. His frequency of use was two or three times per week, on the average, although there were periods of two to three months at a time when he did not use marijuana, and even one period of nine months when he did not smoke marijuana. (Items 3, 6, and 7.) He last smoked marijuana on January 15, 1996, celebrating his "good fortune" in being offered a job with a defense contractor. (Items 3, 6, and 7.) During the period from approximately 1989 to January of 1996, Applicant purchased marijuana approximately ten to twenty times, paying from \$25.00 to \$50.00 per occasion to purchase an eighth of an ounce. (Item 7.)

Applicant arranged and paid for drug screening tests on nine separate dates during a three month period from February 26, 1996 to May 29, 1996, and the results of those tests were negative. (Item 3.) He also arranged and paid for a drug screening test on September 17, 1996, and the results of that test were negative. (Additional Material submitted by Applicant in response to the FORM.)

Applicant's intention is not to smoke marijuana "at least through (his) period of employment with (a defense contractor) and/or the duration of (his) need to hold a security clearance." (Item 7.) Applicant states that "marijuana is still in (his) environment because (he) continues to associate with musicians and other performers about twice a week ..." (Item 7.)

Mitigation.

Applicant's use of illegal drugs has not resulted in any arrests or in any financial problems. He has never trafficked in, sold, distributed, manufactured, or grown any illegal drugs. (Item 7.)

Applicant graduated in the top ten percent of his high school class, and he graduated from college. (Items 3, 4, and 6.)

Applicant arranged and paid for drug screening tests on nine separate dates during a three month period from February 26, 1996 to May 29, 1996, and the results of those tests were negative. (Item 3.) He also arranged and paid for a drug screening test on September 17, 1996, and the results of that test were negative. (Additional Material submitted by Applicant in response to the FORM.)

B.3.1 Advantages of the Dataset

This dataset has several inherent advantages to it that make it a good potential choice to use as an example:

- The dataset is (from what I could find) completely unused by other scholars in the field. This freshness will allow for original research that has not been previously covered.
- The dataset is based on raw text. While machine learning techniques have long dealt with numerical or categorical data, neural networks have shown great promise in the field of textual analysis.
- The data set has genuine public policy relevance – handling the approval or denial of security clearances is a critical task for the federal government to manage. False positives (providing clearances to those who should be denied) is a substantial national security risk, and false negatives (denying those clearances which should be granted) can potentially ruin the careers of dedicated civil servants.
- My baseline neural network will utilize Google's *Word2Vec* model for converting the textual data into machine-readable numerical information. Word2Vec and models like it have become frequent tools for analyzing textual data with neural networks due to their ability to mathematically describe the relationship between words.

At the same time, the dataset and situation has one key disadvantage: in general (although not entirely), the DoD is focused on issues abroad rather than domestically.

Thus, most ML models developed by DoD would be outside the scope of this study (see *Section 1.6*). A perfect test case for this study would be from an organization with an entirely domestic focus. Nevertheless, since the impact of these security clearance decisions is directly on American citizens that are often employed domestically, I believe that this issue is at least partially mitigated.

B.4 Data Pre-Processing

The data required substantial pre-processing prior to use. After extracting the raw data from the DOHA website, I needed to parse out only the most vital information. In particular, not all sections from each clearance case summary were included. I did not want to include the subjective determinations of the judges themselves; this would allow the model to “cheat” more easily. Rather, I only provided the model with the “objective” sections from each clearance case summary – namely, the sections entitled “Statement of the Case” and “Findings of Fact”. Additionally, I deleted any phrases similar to “clearance is granted” or “clearance is denied” to prevent the model from unfairly using such statements to make a circular determination (i.e. clearance predicted to be granted because it says “clearance is granted”). I also worked to extract the applicant’s gender from each application for later analysis, as well as the decision itself, which weren’t always clear from the text.

B.5 Modeling, Testing, and Validation

The model I chose was a Convolutional Neural Network (CNN). While CNNs are most often used for image data, they have also been shown to be successful for textual analysis as well. Recurrent Neural Networks (RNNs) or their cousin Long Short Term Memory neural networks (LSTMs) are generally more common for use with most kinds of natural language processing (NLP) tasks. The difference between RNNs and CNNs is that “[a]n RNN is trained to recognize patterns across time, while a CNN learns to recognize patterns across space” (Ghelani 2019). While this “space” is commonly associated with image data, there is no reason why it can’t work with textual data as well.

Once I knew I was going to use a CNN, the next step was to design the specific architecture and layers of the model. As (a) this is not a computer science dissertation, (b) this is a fairly basic and straightforward sentence/document classification problem, and (c) I do not claim the same level of expertise as a computer science PhD at constructing such models, I simply used an architecture based off of an extremely well-known (if older) model from previous scholarly literature (Kim 2014) (Kekic 2018). I make no claim that this is the ideal architecture to use, or that there are no superior architectures out there. The technical specifics of my implementation of the model can be seen on my associated GitHub project. TensorFlow (the Python software library used to create and train the model) produced the following summary of the model’s structure:

Layer (type)	Output Shape	Param #	Connected to
input_1 (InputLayer)	(None, 5514)	0	
embedding_1 (Embedding)	(None, 5514, 300)	6000000	input_1[0][0]
reshape_1 (Reshape)	(None, 5514, 300, 1)	0	embedding_1[0][0]
conv2d_1 (Conv2D)	(None, 5512, 1, 100)	90100	reshape_1[0][0]
conv2d_2 (Conv2D)	(None, 5511, 1, 100)	120100	reshape_1[0][0]
conv2d_3 (Conv2D)	(None, 5510, 1, 100)	150100	reshape_1[0][0]
max_pooling2d_1 (MaxPooling2D)	(None, 1, 1, 100)	0	conv2d_1[0][0]
max_pooling2d_2 (MaxPooling2D)	(None, 1, 1, 100)	0	conv2d_2[0][0]
max_pooling2d_3 (MaxPooling2D)	(None, 1, 1, 100)	0	conv2d_3[0][0]
concatenate_1 (Concatenate)	(None, 3, 1, 100)	0	max_pooling2d_1[0][0] max_pooling2d_2[0][0] max_pooling2d_3[0][0]
flatten_1 (Flatten)	(None, 300)	0	concatenate_1[0][0]
dropout_1 (Dropout)	(None, 300)	0	flatten_1[0][0]
dense_1 (Dense)	(None, 2)	602	dropout_1[0][0]
Total params: 6,360,902			
Trainable params: 6,360,902			
Non-trainable params: 0			

Figure 17 - TensorFlow Model Summary

B.5.1 Results of the Model

With about 4,000 examples to train on and about 1,000 examples to test on, the model achieved a high accuracy of about 97.5% on the test data during validation. More specifically, the following confusion matrix was created based on 1,002 elements of test data:

Table 21 - Confusion Matrix

Confusion Matrix		
	Predicted TRUE	Predicted FALSE
Actually TRUE	315 (True Positive)	14 (False Negative)
Actually FALSE	11 (False Positive)	662 (True Negative)

With the confusion matrix in hand (see *Section 4.1* above), recall, precision, F1 Score, markedness, and informedness can all be calculated:

Recall: 0.957

Precision: 0.966

F1 Score: 0.962

Informedness: 0.9411

Markedness: 0.946

From these statistics, we can see that no matter which kind of accuracy measurement is used, they all say the same thing: this is an accurate predictive model.

B.5.2 Additional Validation: Testing Other Models

In addition to the CNN, I also tested four other ML algorithms to see if the CNN's results were unique or if these other models could achieve the same results. Each of these other models were cross-validated five times; in other words, the models were each created five separate times with different data points randomly put into the training dataset and the testing dataset each time. This helped to ensure a more robust model, since it is not dependent on some "lucky" data points getting into the training dataset. The four models included: support vector machines, random forests, logistic regression, and multinomial naïve bayes. The accuracy of the other four models' results is below (LinearSVC refers to the SVM):

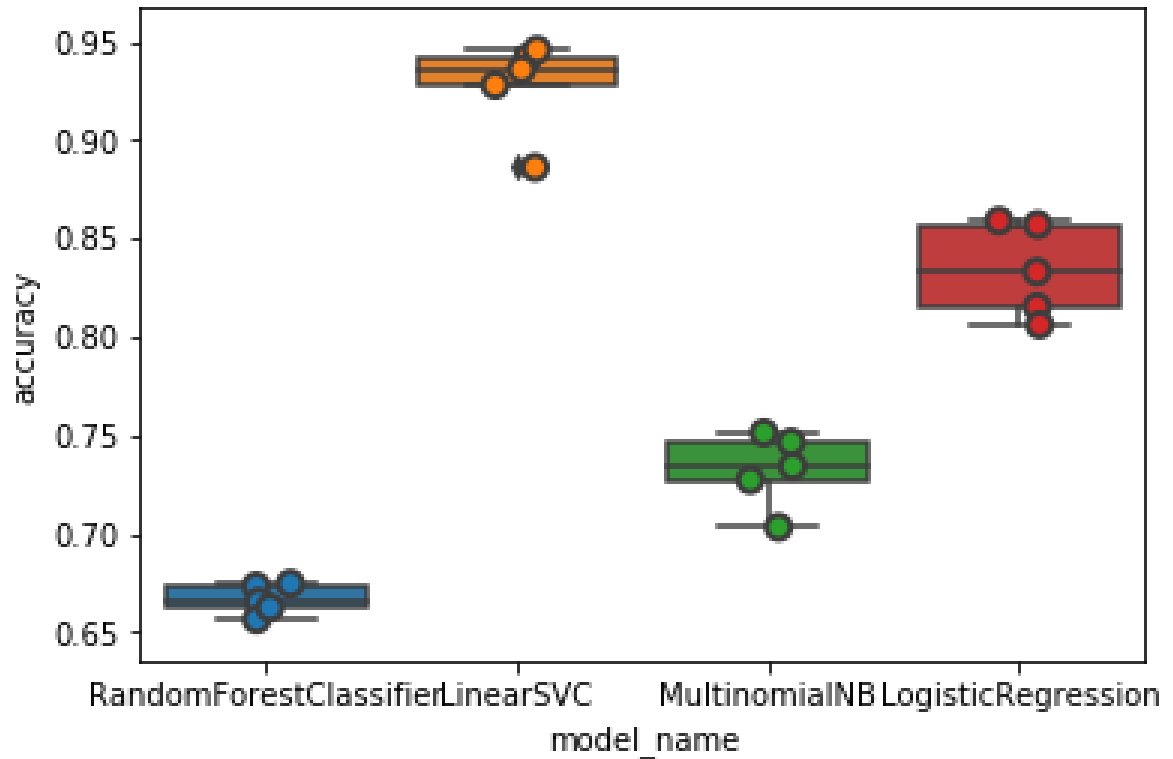


Figure 18 - Comparison to Other ML Models

While the SVM approached the ANN in accuracy (SVMs are often thought to have near-ANN levels of predictive accuracy at many tasks), the other three models clearly did not approach the same level of predictive accuracy. However, they all showed themselves to be relatively robust during cross-validation.

B.5.3 Conclusions from Testing and Validation

At first glance, the CNN model achieves an extremely high level of accuracy and it surpasses a variety of other ML algorithms thrown at the same problem. However, this does not imply that there are no problems to applying such a model in a public

policy setting. *Section 6.6* will delve into the kinds of questions that should arise when attempting to implement such a model.

B.6 Deploy, Monitor, and Reassess

Since this example is not actually being applied in the real world, there is little to add in this section.

Appendix C: Interviewees

The following individuals were interviewed and provided peer reviews to complete Stage Four of this dissertation:

Elizabeth Bondi, PhD Candidate, Harvard University

Robert Brauneis, Professor of Law, George Washington University

Aziz Huq, Frank and Bernice J. Greenberg Professor of Law, University of Chicago

Dr. Daniel Greene, Assistant Professor, University of Maryland

Dr. Gregory Hager, Mandell Bellmore Professor of Computer Science, Johns Hopkins University

References

- Abadicio, Millicent. 2019. *Artificial Intelligence at the FBI – 6 Current Initiatives and Projects*. May 19. <https://emerj.com/ai-sector-overviews/artificial-intelligence-fbi/>.
- Abrash, Victor, Michael Cohen, and Horacio Franco. 1997. *Hybrid Neural Network/Hidden Markov Speech Recognition*. <http://www.speech.sri.com/projects/hybrid.html>.
- Abrassart, Christophe, Yoshua Bengio, Guillaume Chicoisne, Nathalie de Marcellis-Warin, MarcAntoine Dilhac, Sébastien Gambs, and Vincent et al. Gautrais. 2018. *Montréal Declaration for Responsible Development of Artificial Intelligence*. Declaration of Principles, Montréal: Université de Montréal. 2018. Montréal Declaration for Responsible Development of Artificial Intelligence: 1–21.
- Allan, Lorraine G. 1980. "A note on measurement of contingency between two binary variables in judgment tasks ." *Bulletin of the Psychonomic Society* 147-149.
- Allen, Kate. 2015. *How a Toronto professor's research revolutionized artificial intelligence*. April 11. <https://www.thestar.com/news/world/2015/04/17/how-a-toronto-professors-research-revolutionized-artificial-intelligence.html>.
- . 2015. *How a Toronto professor's research revolutionized artificial intelligence*. April 11. <https://www.thestar.com/news/world/2015/04/17/how-a-toronto-professors-research-revolutionized-artificial-intelligence.html>.
- Alshemali, Basemah, and Jugal Kalita. 2019. "Improving the Reliability of Deep Neural Networks in NLP: A Review." *Knowledge-Based Systems*.
- Amnesty International. 2018. *The Toronto Declaration: Protecting the right to equality and non-discrimination in machine learning systems*. Report, Canada: Amnesty International.
- Anastasopoulos, Jason, and Andrew B. Whitford. 2019. "Machine Learning for Public Administration Research, With Application to Organizational Reputation."

Journal of Public Administration Research and Theory, Volume 29, Issue 3 491-510.

Anderson, James A., and Edward Rosenfeld. 1993. *Talking Nets: An Oral History of Neural Networks*. MIT Press.

Andreessen Horowitz. 2017. *Frank Chen will make you a believer in AI*. December 12. <https://mixpanel.com/blog/2017/12/12/frank-chen-ai-andreessen-horowitz/>.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. *Machine Bias*. ProPublica. <https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing>.

Athalye, Anish, Nicholas Carlini, and David Wagner. 2018. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples." *ICML 2018*.

Badr, Will. 2019. *Auto-Encoder: What Is It? And What Is It Used For? (Part 1)*. April 22. <https://towardsdatascience.com/auto-encoder-what-is-it-and-what-is-it-used-for-part-1-3e5c6f017726>.

Bagdasaryan, Eugene, and Vitaly Shmatikov. 2019. "Differential Privacy Has Disparate Impact on Model Accuracy." *ArXiv Pre-Print 1905.12101v1*.

Barker, Jon. 2016. *From the Frontline: How Deep Learning Plays Critical Role in Military Problem-Solving*. June 29. <https://blogs.nvidia.com/blog/2016/06/29/deep-learning-6/>.

Barros, Thiago M., Plácido A. Souza Neto, Ivanovitch Silva, and Luiz Affonso Guedes. 2019. "Predictive Models for Imbalanced Data: A School Dropout Perspective." *Education Sciences*.

Bengio, Yoshua. 1993. "A Connectionist Approach To Speech Recognition." *International Journal of Pattern Recognition and Artificial Intelligence*.

Bergen, Mark, and Kurt Wagner. 2015. *Welcome to the AI Conspiracy: The 'Canadian Mafia' Behind Tech's Latest Craze*. July 15. <https://www.recode.net/2015/7/15/11614684/ai-conspiracy-the-scientists-behind-deep-learning>.

Bojarski, Mariusz, Ben Firner, Beat Flepp, Larry Jackel, Urs Muller, and Karol Zieba. 2016. *End-to-End Deep Learning for Self-Driving Cars*. August 17. <https://devblogs.nvidia.com/parallelforall/deep-learning-self-driving-cars/>.

Bolukbasi, Tolga, Kai-Wei Chang, James Zou, Venkatesh Saligrama, and Adam Kala. 2016. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word

- Embeddings." *30th Conference on Neural Information Processing Systems*. Barcelona, Spain: NIPS 2016.
- Bonnín-Roca, Jaime, Parth Vaishnav, M.Granger Morgan, Joana Mendonça, and Erica Fuchs. 2017. "When risks cannot be seen: Regulating uncertainty in emerging technologies." *Research Policy* 1215-1233.
- Boopathy, Akhilan. 2019. "Presentation on an efficient computation framework of a certified robustness measure for convolutional neural networks." *AAAI*.
- Boopathy, Akhilan, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. 2019. "CNN-Cert: An Efficient Framework for Certifying Robustness of Convolutional Neural Networks." *AAAI*.
- Bourlard, H., and Y. Kamp. 1988. "Auto-association by multilayer perceptrons and singular value decomposition." *Biological Cybernetics* 59 291-294.
- Brajer, Nathan, Brian Cozzi, Michael Gao, Mike Revoir, Marshall Nichols, Joseph Futoma, Jonathan Bae, et al. 2019. "Prospective and External Evaluation of a Machine Learning Model to Predict In-Hospital Mortality." *medRxiv* 19000133; doi: <https://doi.org/10.1101/19000133>.
- Brauneis, Robert, and Ellen P. Goodman. 2018. "Algorithmic Transparency for the Smart City." *Yale Journal of Law and Technology*, Volume 20 103-176.
- Brkan, Maja. 2019. "Do algorithms rule the world? Algorithmic decision-making and data protection in the framework of the GDPR and beyond." *International Journal of Law and Information Technology* 91–121.
- Broderick, James A. 2009. *April 25, 1938 – The Case of Carolene Products, or, The Most Famous Footnote in the History of Law*. April 25. <https://legallegacy.wordpress.com/2009/04/25/april-25-1938-the-case-of-carolene-products-or-the-most-famous-footnote-in-the-history-of-law/>.
- Browlee, Jason. 2017. *A Gentle Introduction to Transfer Learning for Deep Learning*. December 20. <https://machinelearningmastery.com/transfer-learning-for-deep-learning/>.
- Calders, T, and S Verwer. 2010 . "Three naive Bayes approaches for discrimination-free classification." *Data Mining and Knowledge Discovery Volume 21, Issue 2* 277-292.
- Carlini, Nicholas, and David Wagner. 2016. "Defensive Distillation is Not Robust to Adversarial Examples." *arXiv preprint: arXiv:1607.04311v1*.
- . 2017. "Towards Evaluating the Robustness of Neural Networks." *IEEE Symposium on Security and Privacy*. IEEE. 39-57.

- Chang, Hyeon Soo, Michael C. Fu, Jiaqiao Hu, and Steven I. Marcus. 2016. *Operations research's unheralded role in the path-breaking achievement*. October. <https://www.informs.org/ORMS-Today/Public-Articles/October-Volume-43-Number-5/Google-DeepMind-s-AlphaGo>.
- Chen, Irene Y, Fredrik D Johansson, and David Sontag. 2018. "Why Is My Classifier Discriminatory?" *32nd Conference on Neural Information Processing Systems*. Montréal, Canad: NeurIPS 2018.
- Chiocchetti, Paolo. 2017. *Democratic Legitimacy*. August 28. <https://resume.uni.lu/story/democratic-legitimacy>.
- Ciresan, Dan Claudiu, Ueli Meier, Luca Maria Gambardella, and Jurgen Schmidhuber. 2010. "Deep Big Simple Neural Nets Excel on Handwritten Digit Recognition."
- Coglianesi, Cary, and David Lehr. 2017. "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era." *Faculty Scholarship at Penn Law* 1147-1223.
- Coglianesi, Cary, and David Lehr. 2019. "Transparency and Algorithmic Governance." *Administrative Law Review* 1-56.
- Collins, Terry. 2019. *Facial recognition: Do you really control how your face is being used?* November 19. <https://www.usatoday.com/story/tech/2019/11/19/police-technology-and-surveillance-politics-of-facial-recognition/4203720002/>.
- Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. "Algorithmic decision making and the cost of fairness." *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Halifax, NS, Canada: ACM New York. 797-806.
- Cortes, Corinna, and Vladimir Vapnik. 1995. "Support-Vector Networks." *Machine Learning, Volume 20* 273-297.
- Cummings, Rachel, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. 2019. "On the Compatibility of Privacy and Fairness." *ACM UMAP 2019*. Larnaca, Cyprus: ACM.
- Dalvi, Niles, Pedro Domingos, Sumit Sanghai, and Deepak Verma. 2004. "Adversarial Classification." *KDD '04 Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. New York, NY: ACM. 99-108.
- Danaher, John, Michael J. Hogan, Chris Noone, Ronan Kennedy, Anthony Behan, Aisling De Paor, and Heike Felzmann. 2017. "Algorithmic governance: Developing a

- research agenda through the power of collective intelligence." *Big Data & Society* 1-21.
- Data School. 2014. *Simple guide to confusion matrix terminology*. March 25. <https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>.
- Dieterich, William, Christina Mendoza, and Tim Brennan. 2016. *COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity*. Response Study, Northpointe Inc.
- Dorschel, Arianna. 2019. *Rethinking Data Privacy: The Impact of Machine Learning*. April 24. <https://medium.com/luminovo/data-privacy-in-machine-learning-a-technical-deep-dive-f7f0365b1d60>.
- Dowlin, Nathan, Ran Gilad-Bachrach, Kim Laine, Kristin Lauter, Michael Naehrig, and John Wernsing. 2016. "CryptoNets: Applying Neural Networks to Encrypted Data." *Proceedings of the 33rd International Conference on Machine Learning*. New York, NY: JMLR.
- Dryden, Jane. n.d. *Internet Encyclopedia of Philosophy*. Accessed November 11, 2019. <https://www.iep.utm.edu/autonomy/>.
- Dwork, Cynthia. 2006. "Lecture Notes in Computer Science." In *Automata, Languages and Programming*, by M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, 1-12. Springer, Berlin, Heidelberg: ICALP 2006.
- Eagan, Jennifer L. 2013. *Encyclopedia Britannica: Deliberative democracy*. March 12. <https://www.britannica.com/topic/deliberative-democracy>.
- Faes, Livia, Siegfried K. Wagner, Dun Jack Fu, and et al. 2019. "Automated deep learning design for medical image classification by health-care professionals with no coding experience: a feasibility study." *The Lancet Digital Health Volume 1, Issue 5* e232-e242.
- FCC. 2019. *Rulemaking Process*. <https://www.fcc.gov/about-fcc/rulemaking-process>.
- Feldman, Michael, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. "Certifying and removing disparate impact." *Proc. 21st ACM KDD*. 259–268.
- Flores, Anthony W., Kristin Bechtel, and Christopher T. Lowenkamp. 2016. "False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks."." *Federal probation* 80(2).
- FREOPP. 2016. *Our Mission*. June 12. <https://freopp.org/our-mission-3b16e8e8c656>.

- Friedler, Sorelle A., Carlos Scheidegger, and Suresh Venkatasubramanian. 2016. *On the (im)possibility of fairness*. arXiv:1609.07236v1 [cs.CY], arXiv.
- Friedler, Sorelle, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. 2018. "A comparative study of fairness-enhancing interventions in machine learning." *arXiv preprint 1802.04422*.
- Future of Life Institute. 2017. *ASILOMAR AI PRINCIPLES*. Accessed July 30, 2018. <https://futureoflife.org/ai-principles/?submitted=1&cn-reloaded=1#confirmation>.
- Garg, Sahaj, Vincent Perot, Nicole Limtiaco, Ankur Taly, Ed H. Chi, and Alex Beutel. 2019. "Counterfactual Fairness in Text Classification through Robustness." *arXiv prePrint arXiv:1809.10610v2 [cs.LG]*.
- Garson, G. David. 1998. *Neural Networks: An Introductory Guide for Social Scientists*. London: SAGE Publications Ltd.
- Gentry, Craig. 2009. "Fully homomorphic encryption using ideal lattices." *STOC, volume 9* 169 - 178.
- Ghelani, Shreya. 2019. *Text Classification — RNN's or CNN's?* June 1. <https://towardsdatascience.com/text-classification-rnns-or-cnn-s-98c86a0dd361>.
- Gilpin, Leilani H., David Bau, Ben Z. Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. "Explaining Explanations: An Overview of Interpretability of Machine Learning." *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. Turin, Italy: IEEE.
- Globerson, Amir, and Sam Roweis. 2006. "Nightmare at test time: robust learning by feature deletion." *ICML '06 Proceedings of the 23rd international conference on Machine learning*. New York, NY: ACM. 353-360.
- Google. 2019. *Responsible AI Practices*. <https://ai.google/responsibilities/responsible-ai-practices/>.
- Government Accountability Office. 2016. *Face Recognition Technology: FBI Should Better Ensure Privacy and Accuracy*. GAO Report, Washington, D.C.: US Federal Government.
- Green, Matthew. 2016. *What is Differential Privacy?* June 15. <https://blog.cryptographyengineering.com/2016/06/15/what-is-differential-privacy/>.

- Greenemeier, Larry. 2017. *20 Years after Deep Blue: How AI Has Advanced Since Conquering Chess*. June 2. <https://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/>.
- Grun, Felix, Christian Rupprecht, Nassir Navab, and Federico Tombari. 2016. "A Taxonomy and Library for Visualizing Learned Features in Convolutional Neural Networks." *Proceedings of the 33rd International Conference on Machine Learning*. New York, NY: JMLR.
- Guthrie, Susan, Watu Wamae, Stephanie Diepeveen, Steven Wooding, and Jonathan Grant. 2013. *Measuring Research: A guide to research evaluation frameworks and tools*. Washington, DC: RAND Corporation.
- Hagendorff, Thilo. 2019. *The Ethics of AI Ethics: An Evaluation of Guidelines*. Research Paper, Tuebingen: International Center for Ethics in the Sciences and Humanities, University of Tuebingen.
- Hand, David, and Peter Christen. 2018. "A note on using the F-measure for evaluating record linkage algorithms." *Statistics and Computing* 539-547.
- Hardt, Moritz, Eric Price, Nati Srebro, and et al. 2016. "Equality of opportunity in supervised learning." *Advances in neural information processing systems* 3315–3323.
- Harvard. 2019. *Machine Learning for Wildlife Conservation with UAVs*. <https://teamcore.seas.harvard.edu/machine-learning-wildlife-conservation-uavs>.
- Hayes, Adam. 2019. *Harmonic Mean*. October 27. <https://www.investopedia.com/terms/h/harmonicaverage.asp>.
- He, Warren, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. 2017. "Adversarial Example Defenses: Ensembles of Weak Defenses are not Strong." *11th {USENIX} Workshop on Offensive Technologies ({WOOT} 17)*.
- Hebb, Donald. 1949. *Organization of Behavior*. Psychology Press.
- Hendrycks, Dan, and Thomas Dietterich. 2019. "Benchmarking Neural Network Robustness To Common Corruptions and Perturbations." *ICLR*. ICLR. 1-16.
- High-Level Expert Group on AI. 2019. *Ethics Guidelines for Trustworthy AI*. Ethical AI Framework, Brussels: European Commission.
- Hill, Kashmir. 2020. *The Secretive Company That Might End Privacy as We Know It*. January 18. <https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html>.

- Hilligoss, Hannah, and Jessica Fjeld. 2019. *Introducing the Principled Artificial Intelligence Project*. June 7. <https://cyber.harvard.edu/story/2019-06/introducing-principled-artificial-intelligence-project>.
- Hinton, G., O. Vinyals, and J. Anddean. 2015. "Distilling the knowledge in a neural network." *arXiv preprint arXiv:1503.02531*.
- Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A fast learning algorithm for deep belief nets." *Neural Computation*.
- Hinton, Geoffrey E., Simon Osindero, and Yee-Whye Teh. 2006. "A fast learning algorithm for deep belief nets." *Neural Computation*.
- Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. "Long Short-Term Memory." *Neural Computation* 1735-1780.
- Hopfield, J. J. 1982. "Neural networks and physical systems with emergent collective computational abilities." *Proceedings of the National Academy of Sciences April 1, 1982* 79 (8). National Academy of Sciences. 2554-2558.
- Hornik, Kurt, Maxwell Stinchcombe, and Halbert White. 1989. "Multilayer feedforward networks are universal approximators." *Elsevier Volume 2, Issue 5* 359–366.
- Jackson, William. 1998. *DARPA project will study neural network processes*. October 26. <https://gcn.com/articles/1998/10/26/darpa-project-will-study-neural-network-processes.aspx>.
- Jagielski, Matthew, Michael Kearns, Jieming Mao, Alina Oprea, Aaron Roth, Saeed Sharifi-Malvajerdi, and Jonathan Ullman. 2019. "Differentially Private Fair Learning." *arXiv prePrint arXiv:1812.02696v3*.
- Jayaraman, Bargav, and David Evans. 2019. "Evaluating Differentially Private Machine Learning in Practice." *arXiv prePrint arXiv:1902.08874v3 [cs.LG]*.
- Jeeva, Manikandan. 2018. *The Scuffle Between Two Algorithms - Neural Network vs. Support Vector Machines*. September 15. <https://medium.com/analytics-vidhya/the-scuffle-between-two-algorithms-neural-network-vs-support-vector-machine-16abe0eb4181>.
- Johnson, R. Colin. 2009. *Neural nets make a comeback at Darpa*. August 10. http://www.eetimes.com/document.asp?doc_id=1171524.
- Kaggle.com. 2018. *TGS Salt Identification Challenge*. July. <https://www.kaggle.com/c/tgs-salt-identification-challenge>.
- Kamiran, F, and T Calders. 2009. "Classifying without discriminating." *Proc. IC4 09*. IEEE press.

- Katyal, Sonia. 2019. "The Paradox of Source Code Secrecy." *Cornel Law Review*.
- Kehl, Danielle, Priscilla Guo, and Samuel Kessler. 2017. *Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing*. Responsive Communities Initiative, Berkman Klein Center for Internet & Society, Harvard Law School.
- Kekic, Marija. 2018. *CNN in keras with pretrained word2vec weights*.
<https://www.kaggle.com/marijakekic/cnn-in-keras-with-pretrained-word2vec-weights>.
- Kent, Allen, Madeline M. Berry, Fred U. Luehrs Jr., and J. W. Perry. 1955. "Machine literature searching VIII. Operational criteria for designing information retrieval systems." *Journal of the Association for Information Science and Technology*, Volume 6, Issue 2 93-101.
- Kepler, Fábio. 2019. *Why AI fails in the wild*. November 15.
<https://unbabel.com/blog/artificial-intelligence-fails/>.
- Kim, Yook. 2014. "Convolutional Neural Networks for Sentence Classification." *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics. 1746-1751.
- Ko, Tin Kam. 1995. "Random decision forests." *ICDAR '95 Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1)*. Washington, DC: IEEE Computer Society. 278.
- Koene, Ansgar, Chris Clifton, Yohko Hatada, Helena Webb, Menisha Patel, Caio Machado, Jack LaViolette, Rashida Richardson, and Dillon Reisman. 2019. *A governance framework for algorithmic accountability and transparency*. Study for the Panel for the Future of Science and Technology, European Parliamentary Research Service.
- Koulu, Riikka. 2019. "Human Oversight of Automation – Reflections on Ai Ethics, Technological Agency and Anthropocentric Law ." *SSRN* 1-23.
- Kriesi, Hanspeter. 2013. "Democratic legitimacy: Is there a legitimacy crisis in contemporary politics?" *Politische Vierteljahresschrift*, Vol. 54, No. 4 609-638.
- Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. 2012. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25. NIPS.

- Kurenkov, Andrey. 2015. A 'Brief' History of Neural Nets and Deep Learning, Part 2. December 24. <http://www.andreykurenkov.com/writing/ai/a-brief-history-of-neural-nets-and-deep-learning-part-2/>.
- Laskai, Lorand, and Graham Webster. 2019. *Translation: Chinese Expert Group Offers 'Governance Principles' for 'Responsible AI'*. June 17. <https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/>.
- LeCun, Y, B Boser, J Denker, D Henderson, R Howard, W Hubbard, and L Jackel. 1989. "Backpropagation Applied to Handwritten Zip Code Recognition." *Neural Computation*, vol.1, no.4 541-551.
- LeCun, Y., L. D. Jackel, L. Bottou, A. Brunot, C. Cortes, J. S. Denker, H. Drucker, et al. 1995. "Comparison of learning algorithms for handwritten digit recognition." *International Conference on Artificial Neural Networks*. Paris. 53-60.
- Lecuyer, Mathias, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. "Certified Robustness to Adversarial Examples with Differential Privacy." *arXiv:1802.03471v4 [stat.ML]*.
- Lecuyer, Mathias, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2018. "On the Connection between Differential Privacy and Adversarial Robustness in Machine Learning." *arXiv prePrint arXiv:1802.03471v1*.
- Lee, Wei-Han, Changchang Liu, Shouling Ji, Prateek Mittal, and Ruby B. Lee. 2017. *WPES'17*. Dallas, TX, USA.: Association for Computing Machinery.
- Legendre, Adrien-Marie. 1805. "Sur la Méthode des moindres quarrés." In *Nouvelles méthodes pour la détermination des orbites des comètes*, by Adrien-Marie Legendre. Firmin Didot, Paris.
- Lei, Suhua, Zhang, Huan, Ke Wang, and Zhendong Su. 2019. "How Training Data Affect the Accuracy and Robustness of Neural Networks for Image Classification." *ICLR 2019 (under review)*. ICLR.
- Leslie, David. 2019. *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. Framework, London, United Kingdom: Alan Turing Institute.
- Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *2016 ICML Workshop on Human Interpretability in Machine Learning*. New York, NY, USA: WHI.
- Lovejoy, Josh. 2018. *The UX of AI*. Report, Google.

- Madry, Aleksander, Aleksander Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2019. "Towards Deep Learning Models Resistant to Adversarial Attacks." *arXiv:1706.06083v4*.
- Mancuso, Jason. 2019. *Privacy-Preserving Machine Learning 2018: A Year in Review*. January 10. <https://medium.com/dropoutlabs/privacy-preserving-machine-learning-2018-a-year-in-review-b6345a95ae0f>.
- Mangal, Ravi, Aditya Nori, and Alessandro Orso. 2019. "Robustness of Neural Networks: A Probabilistic and Practical Perspective." *ICSE*. Montreal, Canada.
- Martschenko, Daphne. 2017. *Can an IQ Test Really Measure Your Intelligence?* October 11. <https://psmag.com/education/what-do-iq-tests-measure>.
- Matthews, Dylan. 2015. *The case against equality of opportunity*. September 21. <https://www.vox.com/2015/9/21/9334215/equality-of-opportunity>.
- McCarthy, John, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. 1955. "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence." Research Project Proposal.
- McCulloch, Warren S., and Walter H. Pitts. 1943. "A logical calculus of the ideas immanent in nervous activity." *Bulletin of Mathematical Biophysics* 115-133.
- McLeod, Saul. 2017. *The Milgram Shock Experiment*. <https://www.simplypsychology.org/milgram.html>.
- McMahan, Brendan, and Daniel Ramage. 2017. *Federated Learning: Collaborative Machine Learning without Centralized Training Data*. April 6. <https://ai.googleblog.com/2017/04/federated-learning-collaborative.html>.
- Merriam-Webster Dictionary. n.d. *Accuracy*. Accessed October 10, 2019. <https://www.merriam-webster.com/dictionary/accuracy>.
- Mikulski, Bartosz. 2019. *F1 score explained*. February 4. <https://www.mikulskibartosz.name/f1-score-explained/>.
- Minsky, Marvin, and Seymour A. Papert. 1969. *Perceptrons*. MIT Press.
- Mitsa, Theophano. 2019. *How Do You Know You Have Enough Training Data?* April 22. <https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee>.
- Mittal, Aditi. 2019. *Understanding RNN and LSTM*. October 12. <https://towardsdatascience.com/understanding-rnn-and-lstm-f7cdf6dfc14e>.

- Mittelstadt, Brent Daniel, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. "The ethics of algorithms: Mapping the debate." *Big Data & Society* 1–21.
- Mohamed, Abdel-rahman, George Dahl, and Geoffrey Hinton. 2009. "Deep Belief Networks for phone recognition."
- Moisejevs, Ilja. 2019. *Poisoning attacks on Machine Learning*. July 14. <https://towardsdatascience.com/poisoning-attacks-on-machine-learning-1ff247c254db>.
- Narendra, K. S., and K Parthasarathy. 1990. "Identification and control of dynamical systems using neural networks." *Neural Networks, IEEE Transactions* 4-27.
- New York Times. 1958. *NEW NAVY DEVICE LEARNS BY DOING; Psychologist Shows Embryo of Computer Designed to Read and Grow Wiser*. July 8. <http://www.nytimes.com/1958/07/08/archives/new-navy-device-learns-by-doing-psychologist-shows-embryo-of.html>.
- Olazaran, Mikel. 1996. "A Sociological Study of the Official History of the Perceptrons Controversy." *Social Studies of Science Vol. 26, No. 3* 611-659.
- Papernot, Nicolas, and Patrick McDaniel. 2017. "Extending Defensive Distillation." *arXiv preprint: arXiv:1705.05264v1*.
- Papernot, Nicolas, Patrick McDaneil, Xi Wu, Somesh Jha, and Ananthram Swami. 2016. "Distillation as a Defense to Adversarial Perturbations against Deep Neural Networks." *37th IEEE Symposium on Security & Privacy*. San Jose, CA: IEEE.
- Papineni, Kishore, Salim Roukous, Todd Ward, and Wei-Jing Zhu. 2002. "BLEU: a Method for Automatic Evaluation of Machine Translation." *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia: ACL. 311-318.
- Pásztor, Dávid. 2018. *AI UX: 7 Principles of Designing Good AI Products*. April 17. <https://uxstudioteam.com/ux-blog/ai-ux/>.
- Pedreschi, Dino, Salvatore Ruggieri, and Franco Turini. 2009. "Measuring Discrimination in Socially-Sensitive Decision Records." *Proceedings*. Sparks, Nevada: 2009 Siam International Conference on Data Mining.
- Peters, Anne. 2011. "The Subjective International Right." *Jahrbuch des öffentlichen Rechts der Gegenwart, Vol. 59* 411-456.
- Phan, NhatHai, Minh Vu, Yang Liu, Ruoming Jin, Dejing Dou, Xintao Wu, and My T. Thai. 2019. "Heterogeneous Gaussian Mechanism: Preserving Differential Privacy in

- Deep Learning with Provable Robustness." *arXiv prePrint arXiv:1906.01444v1 [cs.CR]*.
- Phan, NhatHai, My T. Thai, Ruoming Jin, Han Hu, and Dejing Dou. 2019. "Preserving Differential Privacy in Adversarial Learning with Provable Robustness." *arXiv prePrint arXiv:1903.09822v2 [cs.CR]*.
- Pichai, Sundar. 2018. *AI at Google: our principles*. June 7. <https://www.blog.google/technology/ai/ai-principles/>.
- Powers, David. 2011. "EVALUATION: FROM PRECISION, RECALL AND F-MEASURE TO ROC, INFORMEDNESS, MARKEDNESS & CORRELATION ." *Journal of Machine Learning Technologies* 37-63.
- Price II, W. Nicholson. 2017. "Regulating Black-Box Medicine." *Michigan Law Review Volume 116, Issue 3* 421-474.
- Raff, Edward, and Jared Sylvester. 2018. "Gradient Reversal against Discrimination: A Fair Neural Network Learning Approach." *IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. Turin, Italy: IEEE.
- Raghupathi, Viju, and Wullianallur Raghupathi. 2017. "Preventive Healthcare: A Neural Network Analysis of Behavioral Habits and Chronic Diseases." *Healthcare*.
- Ram, Natalie. 2017. "Innovating Criminal Justice." *Northwestern University Law Review*.
- Raschka, Sebastian. 2019. *Machine Learning FAQ: What is the relation between Logistic Regression and Neural Networks and when to use which?* <https://sebastianraschka.com/faq/docs/logisticregr-neuralnet.html>.
- Raso, Filippo, Hannah Hilligoss, Vivek Krishnamurthy, Christopher Bavitz, and Levin Kim. 2018. *Artificial Intelligence & Human Rights: Opportunities & Risks*. Center Report, Boston, MA: Berkman Klein Center for Internet and Society at Harvard University.
- Ray, Shaan. 2018. *History of AI*. August 11. <https://towardsdatascience.com/history-of-ai-484a86fc16ef>.
- Ray, Sunil. 2017. *Commonly used Machine Learning Algorithms (with Python and R Codes)*. September 9. <https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>.
- Ray, Tiernan. 2019. *Facebook's latest giant language AI hits computing wall at 500 Nvidia GPUs*. November 12. <https://www.zdnet.com/article/facebooks-latest-giant-language-ai-hits-computing-wall-at-500-nvidia-gpus/>.

- Redden, Joanna. 2018. "Democratic governance in an age of datafication: Lessons from mapping government discourses and practices." *Big Data & Society*.
- Reisman, Dillon, Jason Schultz, Kate Crawford, and Meredith Whittaker. 2018. *Algorithmic Impact Assessments: A Practical Framework for Public Agency Accountability*. Study, New York City: AI Now Institute.
- Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier." *KDD '16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco, CA: ACM. 1135-1144 .
- Rivest, Ronald L., Len Adleman, and Michael L. Dertouzos. 1978. "On data banks and privacy homomorphisms." *Foundations of Secure Computation* 4(11) 169 - 180.
- Roberts, Claudia Veronica. 2018. *QUANTIFYING THE EXTENT TO WHICH POPULAR PRE-TRAINED CONVOLUTIONAL NEURAL NETWORKS IMPLICITLY LEARN HIGH-LEVEL PROTECTED ATTRIBUTES*. Master's Thesis, Princeton University.
- Rodriguez, Jesus. 2018. *Understanding Hyperparameters Optimization in Deep Learning Models: Concepts and Tools*. August 8.
<https://towardsdatascience.com/understanding-hyperparameters-optimization-in-deep-learning-models-concepts-and-tools-357002a3338a>.
- Rosenblat, Alex, Tamara Kneese, and Danah Boyd. 2014. *Interpretation Gone Wrong: The Social, Cultural & Ethical Dimensions of "Big Data"*. Workshop Primer, New York City, NY: Data & Society.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning internal representations by error propagation." In *Parallel distributed processing: explorations in the microstructure of cognition*, by David E. Rumelhart, James L. McClelland and PDP Research Group, 318-362. Cambridge, MA: MIT Press.
- Rumelhart, David E., Geoffrey E. Hinton, and Ronald J. Williams. 1986. "Learning Representations by Back-Propagating Errors." *Nature* 533-536.
- Russo, Cami. 2018. *The Human Bias in the AI Machine: How artificial intelligence is subject to cognitive bias*. February 6.
<https://www.psychologytoday.com/ca/blog/the-future-brain>.
- Samuel, Arthur L. 1959. "Some Studies in Machine Learning Using the Game of Checkers." *IBM Journal*, Vol. 3, No.3 535-554.
- Selbst, Andrew D., Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." *FAT**

'19: *Conference on Fairness, Accountability, and Transparency*. Atlanta, GA, USA: 2019 Association for Computing Machinery.

Sharma, Shubham, Jette Henderson, and Joydeep Ghosh. 2019. "CERTIFAI: Counterfactual Explanations for Robustness, Transparency, Interpretability, and Fairness of Artificial Intelligence models." *arXiv prePrint: 1905.07857v1*.

Shokri, Reza, and Vitaly Shmatikov. 2015. "Privacy-Preserving Deep Learning." *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*. Denver, Colorado: ACM. 1310-1321.

Shrum, Karen, Lisa Gordon, Priscilla Regan, Karl Maschino, Alan R. Shark, and Anders Shropshire. 2019. *AI and Its Impact on Public Administration*. Standing Committee Report, National Academy of Public Administration.

Shung, Koo Ping. 2018. *Accuracy, Precision, Recall or F1?* March 15.
<https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9>.

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. 2013. "Deep inside convolutional networks: Visualising image classification models and saliency maps." *CoRR*.

Sinders, Caroline. 2018. *Why the Government Sucks at Making Websites*. June 20.
<https://gizmodo.com/why-the-government-sucks-at-making-websites-1826769004>.

Singh, Gagandeep, Timon Gehr, Markus Puschel, and Martin Vechev. 2019. "Boosting Robustness Certification of Neural Networks." *ICLR*.

Singh, Seema. 2018. *Why correlation does not imply causation?* August 24.
<https://towardsdatascience.com/why-correlation-does-not-imply-causation-5b99790df07e>.

Skitka, Linda. 2011. <https://lskitka.people.uic.edu/styled-7/styled-14/>.

SRI. 1997. *References*.
<http://www.speech.sri.com/projects/hybrid/publications/node1.html>.

Stackify. 2017. *What is SDLC? Understand the Software Development Life Cycle*. April 6.
<https://stackify.com/what-is-sdlc/>.

Statt, Nick. 2019. *OpenAI's Dota 2 AI steamrolls world champion e-sports team with back-to-back victories*. April 13.
<https://www.theverge.com/2019/4/13/18309459/openai-five-dota-2-finals-ai-bot-competition-og-e-sports-the-international-champion>.

- Stoyanovich, Julia. 2019. "TransFAT: Translating Fairness, Accountability and Transparency into Data Science Practice." *PIE*.
- Su, Dong, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. 2019. "Is Robustness the Cost of Accuracy? – A Comprehensive Study on the Robustness of 18 Deep Image Classification Models." *arXiv prePrint 1808.01688v2*.
- Sunstein, Cass. 2015. "The Ethics of Nudging." *Yale Journal on Regulation, Volume 32, Issue 2*.
- SyncedReview. 2019. *Stanford Open-Sources Neural Network Verification Project*. January 2. <https://medium.com/syncedreview/stanford-open-sources-neural-network-verification-project-d2dba0cd21f6>.
- Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. "Intriguing properties of neural networks." *International Conference on Learning Representations (ICLR) 2014*. Banff, Canada: ICLR.
- Thaler, Richard H, and Cass B. Sunstein. 2008. *Nudge: Improving Decisions about Health, Wealth, and Happiness*. Yale University Press.
- The Public Voice. 2018. *Universal Guidelines for Artificial Intelligence*. Report, Brussels, Belgium: The Public Voice.
- Trump, Donald. 2019. *Executive Order on Maintaining American Leadership in Artificial Intelligence*. February 11. <https://www.whitehouse.gov/presidential-actions/executive-order-maintaining-american-leadership-artificial-intelligence/>.
- Tsipras, Dimitris, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Mądry. 2018. "Robustness May Be at Odds with Accuracy." *arXiv prePrint 1805.12152v3*.
- UNI Global Union. 2017. *TOP 10 PRINCIPLES FOR ETHICAL ARTIFICIAL INTELLIGENCE*. Report, Switzerland: UNI Global Union.
- Vahdat, Arash. 2017. "Toward Robustness against Label Noise in Training Deep Discriminative Neural Networks." *31st Conference on Neural Information Processing Systems*. Long Beach, CA, USA: NIPS 2017.
- Valentine, Sarah. 2019. "IMPOVERISHED ALGORITHMS: MISGUIDED GOVERNMENTS, FLAWED TECHNOLOGIES, AND SOCIAL CONTROL." *Fordham Urban Law Journal* 364-427.
- Vigen, Tyler. 2015. *Spurious Correlations*. May 12. <https://www.tylervigen.com/spurious-correlations>.

- Wadsworth, Christina, Francesca Vera, and Chris. Piech. 2018. "Achieving Fairness through Adversarial Learning: an Application to Recidivism Prediction." *ArXiv Pre-Print*.
- Weber, Patrick, Nicolas Weber, Michael Goesele, and Rüdiger Kabst. 2017. "Prospect for Knowledge in Survey Data: An Artificial Neural Network Sensitivity Analysis." *Social Science Computer Review*.
- Weng, Tsui-Wei, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S. Dhillon, and Luca Daniel. 2018. "Towards Fast Computation of Certified Robustness for ReLU Networks." *Proceedings of the 35th International Conference on Machine Learning*. Stockholm, Sweden: PMLR 80, 201.
- Weng, Tsui-Wei, Pin-Yu Chen, Lam M. Nguyen, Mark S. Squillante, Akhilan Boopathy, Ivan Oseledets, and Luca Daniel. 2019. "PROVEN: Verifying Robustness of Neural Networks with a Probabilistic Approach." *Proceedings of the 36th International Conference on Machine Learning*. Long Beach, California: PMLR 97, 2019.
- Werbos, Paul. 1974. *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD Thesis, Cambridge, MA: Harvard University.
- Wexler, Rebecca. 2017. "Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System." *Stanford Law Review*.
- Whittlestone, Jess, Rune Nyrop, Anna Alexandrova, and Stephen Cave. 2019. "The Role and Limits of Principles in AI Ethics: Towards a Focus on Tensions." *Association for the Advancement of Artificial*. Cambridge: Leverhulme Centre for the Future of Intelligence, University of Cambridge.
- Widrow, Bernard. 1989. *DARPA Neural Network Study*. Technical Report, Washington: DARPA.
- Widrow, Bernard, and Marcian Hoff. 1960. *An Adaptive "Adaline" neuron using chemical "memistors"*. Technical Report No. 1553-2, Stanford: Office of Naval Research.
- XenonStack. 2017. *Overview of Artificial Neural Networks and its Applications*. July 17. <https://hackernoon.com/overview-of-artificial-neural-networks-and-its-applications-2525c1addff7>.
- York, Travis T., Charles Gibson, and Susan Rankin. 2015. "Defining and Measuring Academic Success." *Practical Assessment, Research, and Evaluation, Volume 20, Number 5* 1-20.
- Young, Meg, Luke Rodriguez, Emily Keller, Feiyang Sun, Boyang Sa, Jan Whittington, and Bill Howe. 2018. "Beyond Open vs. Closed: Balancing Individual Privacy and

- Public Accountability in Data Sharing." *Proceedings of ACM (FAT '19)*. New York: ACM.
- Yu, Bowen, Ye Yuan, Loren Terveen, Zhiwei Steven Wu, and Haiyi Zhu. 2019. "Designing Interfaces to Help Stakeholders Comprehend, Navigate, and Manage Algorithmic Trade-Offs." *arXiv preprint: arXiv:1910.03061v2 [cs.HC]* 23 Oct 2019.
- Yu, Lei, Ling Liu, Calton Pu, Mehmet Emre Gursoy, and Stacey Truex. 2019. "Differentially Private Model Publishing for Deep Learning." *Proceedings of the 40th IEEE Symposium on Security and Privacy*. Oakland: arXiv:1904.02200 [cs.CR].
- Yurochkin, Mikhail, and Amanda Bower. 2019. "LEARNING FAIR PREDICTORS WITH SENSITIVE SUBSPACE ROBUSTNESS." *arXiv:1907.00020v1 [stat.ML]*.
- Zeiler, Matthew D., and Rob. Fergus. 2013. "Visualizing and understanding convolutional networks." *CoRR*. <http://arxiv.org/abs/1311.2901>.
- Zhang, Hongyang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. 2019. "Theoretically Principled Trade-off between Robustness and Accuracy." *ICML 2019*. ICML.
- Zhang, Huan, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. 2018. "Efficient Neural Network Robustness Certification with General Activation Functions." *32nd Conference on Neural Information Processing Systems*. Montréal, Cana: NeurIPS.
- Zhong, Ziyuan. 2018. *A Tutorial on Fairness in Machine Learning*. October 21. <https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb>.
- Zhou, Victor. 2019. *An Introduction to Recurrent Neural Networks for Beginners*. July 25. <https://towardsdatascience.com/an-introduction-to-recurrent-neural-networks-for-beginners-664d717adbd>.

Biography

Joshua A. Lee graduated from Miami Country Day School in Miami, Florida in 2005. He received his Bachelor of Arts from the University of Central Florida in 2010, his Master of Arts from American University in 2013, and a Graduate Certificate in Cybersecurity Technology from University of Maryland Global Campus in 2016. His dissertation topic was inspired by the 2016 defeat of professional Go player Lee Sedol by the AlphaGo AI. When not dissertating, Josh enjoys tennis, skiing, video games, and the ancient board game of Go.