FREE, OPEN-SOURCE, AND ANONYMOUS: WHY DEEP LEARNING REGULATORS ARE IN DEEP WATER

A THESIS

SUBMITTED TO THE

INTERSCHOOL HONORS PROGRAM IN INTERNATIONAL SECURITY STUDIES

Center for International Security and Cooperation

Freeman Spogli Institute for International Studies

STANFORD UNIVERSITY

By
Andrew Milich

May 2019

Adviser:
Amy Zegart

Acknowledgments

I am deeply grateful to my advisor, Professor Amy Zegart, for her invaluable guidance and instruction over the last two years. Professor Martha Crenshaw, Professor Coit Blacker, and Dr. Kerry Persen also provided incredibly instructive input during my writing process. Admiral James Ellis, Dr. Siegfried Hecker, and Dr. Rick Zhang supplied important insights and research directions as well. Finally, I am thankful for my fellow CISAC honors students' advice and encouragement and for my family's unwavering support.

Contents

List of Tables

List of Figures

Abstract

Deep learning models have been instrumental in driving recent breakthroughs in artificial intelligence. Beyond autonomous navigation and game-playing, some deep learning applications, such as facial recognition and deep fakes, or counterfeit audio and video created by algorithms, pose challenges to individual privacy and US national security. The Director of National Intelligence's 2019 Worldwide Threat Assessment explicitly mentions the growing threat from deep fakes and machine learning systems. Yet, as most deep learning software, datasets, and academic papers are publicly accessible, individuals can effortlessly access the technical prerequisites required to develop deep learning models, build surveillance systems, and access instruments of mass deception. Thousands of deep fakes of politicians and celebrities have already been shared on the internet. The first step towards combating this threat is understanding how deep learning spreads and how its proliferation process differs from other dual-use technologies, or technologies with both civilian and military applications. This thesis seeks to fill this crucial gap by examining how all three critical components of deep learning - data, software, and hardware - have become accessible to internet users. We analyze how existing approaches to mitigating risks from dual-use technology, including establishing norms, limiting supply, and controlling exports, may fail to effectively delay or prevent deep learning threats. After examining how deep fake technology spread from academia to individuals, we present an original experimental study of how technical countermeasures could confuse a facial recognition deep learning model and be used to mitigate risks from deep fakes. Our experiments indicate how US policymakers could leverage both technical and policy mechanisms to delay or undermine malicious deep learning systems.

Chapter 1

Introduction: Deep learning and AI

1.1 AI and societal impact

In the early 2000s, cutting-edge AI could play chess against humans,[1] transcribe speech into text,[2] and drive the first generation of autonomous cars.[3] Although these applications suggested a promising future for AI's role in society, algorithms still struggled to identify faces or objects in images,[4] control complex robots, or synthesize original content.

Today, the latest AI algorithms can automatically colorize black and white photos,[5] read lips from video recordings,[6] and beat human experts at the ancient board game Go.[7] Computer vision algorithms have outperformed doctors at identifying disease markers in thousands of chest x-rays.[8] Identifying objects in pictures - such as animals, cars, buildings, or road signs - is now a routine function that can be performed (relatively) reliably by free software or internet queries.[9] Other models can paint pictures in the style of famous artists,[10] compose classical music,[11] perform realistic human gestures,[12] or generate counterfeit audio and video.[13] Beyond art and music, this new

1. "IBM100 - Deep Blue" 2019.
2. "BBC - iWonder - AI: 15 key moments in the story of artificial intelligence" 2019.
3. "Google's self-driving-car project becomes a separate company: Waymo - Los Angeles Times" 2016.
4. Simonite 2014.
5. Zhang, Isola, and Efros 2016.
6. "astorfi/lip-reading-deeplearning: Lip Reading - Cross Audio-Visual Recognition using 3D Architectures" 2019.
7. "AlphaGo | DeepMind" 2019.
8. Rajpurkar et al. 2017.
9. "Vision API - Image Content Analysis | Cloud Vision API | Google Cloud" 2019.
10. "Artistic Style Transfer with Convolutional Neural Network" 2019.
11. Colombo and Gerstner 2018.
12. Knight 2018.
13. Liu, Breuel, and Kautz 2017.

generation of AI has prompted individuals, states, and companies to pioneer new mass surveillance technology, experiment with autonomous weapons, and create realistic fake videos of politicians and celebrities.[14] In driving many of these recent applications, deep learning has dramatically transformed the research areas, applications, and societal consequences of artificial intelligence (AI).

1.2 Deep learning's role in AI breakthroughs

Sophisticated algorithms, fast hardware, and large datasets have catalyzed recent progress in AI.[15] In particular, research interest in deep learning - a subfield of machine learning - has surged. As deep learning research has precipitated progress in creating deep fakes (falsified audio or video generated by algorithms), identifying faces in images, and driving autonomous vehicles, lawmakers and technologists have recognized that deep learning constitutes an emerging technology with immense societal impact.[16] Furthermore, beyond arcane academic papers, this technology can now be exploited by individual internet users with no programming experience, who can create deep fakes or identify faces in photos without technical skills.[17] Thus, as deep learning models and research are made accessible, the technology's applications will become increasingly difficult to manage, predict, or restrict. Given this democratization of deep learning capabilities, this thesis answers the following question:

How can individuals and governments mitigate the potential abuse of deep learning?

1.3 Argument

No published research has synthesized recent academic papers and open-source software development to trace deep learning proliferation or analyze its national security risks. Although some deep learning experts are aware of the technology's harmful applications, few have considered regulations or methods for limiting malicious uses. Similarly, while policymakers have expressed rising concerns over deep learning and its applications, many lack familiarity with the key technical components driving proliferation and widespread use. As a result, it is critical that regulators and

14. Vincent 2018a.
15. "Stanford University CS231n: Convolutional Neural Networks for Visual Recognition" 2019.
16. Castelvecchi 2019.
17. "FakeApp download links and How-To Guide : GifFakes" 2019.

technologists understand how harmful applications of deep learning spread from academic research to open-source software. This thesis provides technical and analytical evidence to support three hypotheses:

1. Deep learning's technological components yield a unique proliferation method;
2. Existing policy approaches to limiting the spread of other dual-use technology, such as fissile material and rocket engines, are likely to be ineffective in preventing malicious deep learning applications;
3. Technical countermeasures could effectively mitigate risk by delaying or disrupting deep learning systems.

1.4 Outline

Chapter 2 begins by introducing deep learning's historical and theoretical foundations. In the 1950s, computer science and neuroscience researchers began to use software and electrical components to model how information flows through biological neurons. Networks of artificial neurons - called neural networks - demonstrated increasingly promising results across a variety of machine learning applications, such as recognizing letters and digits. In the 1990s and 2000s, advances in computing power and larger datasets led to significant progress in designing and applying artificial neural networks (now called "deep" neural networks) for more complex tasks, including facial and object recognition. After describing how a simple neural network functions, we split deep learning into its three essential enabling components - data, software, and hardware - to analyze when, how, and how fast each component became widely accessible to researchers and individuals.

Chapter 3 analyzes existing methods for controlling the spread of dual-use technology. We begin by examining how technology diffuses through society and how policymakers measure the social or environmental impact of new technology. We subsequently divide past approaches for limiting proliferation into three categories: Norms and taboos, limiting supply, and controlling exports. In each case, we analyze how existing methods for limiting proliferation, which have been applied to restrict the spread of nuclear weapons, biological agents, and military technology, are inadequate for governing deep learning. In particular, deep learning's accessibility to anonymous internet users

poses unique challenges for tracking and enforcement.

Chapter 4 supports our hypothesis about deep learning's unique proliferation process by examining the case of deep fakes. In 2016, deep fake technology was first showcased by academic and commercial research laboratories. Over the next two years, the technology was open-sourced and repackaged as accessible tools for individuals with no coding experience. We analyze how anonymous developers relied on open-source software to transform academic models into easy-to-use applications. This case highlights how existing policies for governing dual-use technology, including norms, limiting supply, and controlling exports, may be unable to prevent deep fake proliferation and limit national security consequences.

Having concluded that existing methods are unlikely to effectively prevent harmful deep learning applications, Chapter 5 examines how technical countermeasures could mitigate deep learning risks. Technical countermeasures are technology developed to prevent, delay, or undermine harmful deep learning applications. Chapter 5 begins by describing how adversarial learning techniques have been developed to confuse or trick deep learning systems. Then, in order to evaluate how technical countermeasures could supplement existing approaches, we performed a set of experiments to attack a deep learning facial recognition model. Our experiments demonstrated how relatively simple attacks could significantly reduce the accuracy and confidence of a facial recognition model. Because facial recognition is required to create deep fakes, these technical mechanisms could be used to limit deep fake creation or be applied to delay or disrupt other deep learning applications.

As research and open-source software development have become geographically decentralized, deep learning cannot be metaphorically "put back in the bottle." Strict export controls and regulatory regimes, as applied to rocket engines, nuclear material, and chemical agents, cannot comprehensively address the national security risks from this technology. The experiments and research in this thesis reveal another approach: Instead of completely restricting deep learning's spread or use, policymakers could reduce risk by tracking malicious actors, delaying harmful applications, and developing technical countermeasures for disrupting deep learning systems. This thesis concludes by proposing a variety of technical and policy mechanisms that could accomplish these objectives.

Chapter 2

Deep learning: Now diffusing rapidly

2.1 Deep fakes and malicious deep learning

Consider the implications of fake audio recordings posted on Twitter or Facebook days before an election, or counterfeit videos of the President making a statement about an impending terrorist attack or public health crisis. On a more personal level: What if an anonymous internet user shared a fake video of you or a family member in humiliating or offensive circumstances? These situations are not hypothetical thought experiments: Open-source deep fake applications released in the last two years have endowed internet users with the power to create realistic-looking but completely fake audio and video.

Deep fakes - or counterfeit audio and video generated by deep learning algorithms - could change the outcome of an election, undermine public confidence in elected officials, or confuse critical decision-making processes in the government or military. Individuals could create supercharged "fake news," release videos of world leaders making falsified statements, or disseminate fake audio recordings to distract attention from genuine offensive content. Given the severity of this threat, a bipartisan group of elected officials, including Republican Senator Marco Rubio and Democratic Congressman Adam Schiff, have warned of deep fakes' pressing national security risks.[1]

News articles and think tank reports echo lawmakers growing concerns. Notable headlines include "Will Deep-Fake Technology Destroy Democracy?" (a *New York Times* opinion article),[2]

1. "2018-09 ODNI Deep Fakes letter.pdf" 2019.
2. Boylan 2018.

"You thought fake news was bad? Deep fakes are where truth goes to die" (an article in *The Guardian*),[3] and "Artificial intelligence, deepfakes, and the uncertain future of truth" (a *Brookings* study).[4] In addition to their impact on democracy and elections, deep fakes pose enormous consequences for individuals; thousands of misleading or compromising deep fakes of celebrities and other individuals have been released on the internet.[5] As these videos can be created and distributed anonymously, it is challenging or impossible to comprehensively remove deep fakes from the internet.[6]

In 2016 - only three years ago - the technology underlying deep fakes was first showcased in academic papers, conference presentations, and industry prototypes. Researchers imagined that their work could enhance teleconferencing, video editing, or special effects.[7] Yet, in only two years, this technology had been repackaged into freely accessible tools for generating fake videos and disseminating disinformation. Deep fake technology has become accessible to individuals with no technical background in machine learning or programming. Without writing a single line of code, an anonymous internet user can easily download an application,[8] select photographs of an individual, and generate fake videos. The application automatically manages the entire process of detecting and transplanting faces.

How does deep learning spread from academia to individuals?

Understanding deep learning proliferation

Tracing how deep learning spreads represents the first step in addressing its growing national security consequences. This chapter answers three questions fundamental to understanding deep learning.

1. What is deep learning?
2. What are the different technological components of deep learning?
3. How do these components proliferate?

3. Schwartz 2018.
4. Villasenor 2019.
5. D. Lee 2018b.
6. D. Lee 2018a.
7. Thies et al. 2016.
8. FakeApp is one example of an application designed for creating deep fakes.

We begin by analyzing the evolution, functionality, and technological components of deep learning models. This chapter provides a brief history of deep learning, which initially piqued researchers' attention in the 1950s when mathematical models were created to approximate how human neurons process information. Inspired by biological processes, researchers applied interconnected architectures of neurons - i.e. neural networks - to machine learning tasks. Over the following half-century, deep learning evolved into a distinct subset of artificial intelligence as neural networks demonstrated increasingly promising results across a variety of machine learning applications, such as object recognition. In order to contextualize these developments, we describe how a single neuron transforms input data and how more complex neural networks are trained to classify or generate data. We also examine how specific types of deep learning models have been applied to facial recognition and creating deep fakes, which present particularly significant implications for national security.

The remainder of this chapter establishes how deep learning models proliferate. We divide deep learning models into three essential technical components: Data, software, and hardware. Software algorithms use data in order to train and test deep learning models; computing hardware runs these computationally intensive algorithms. For all three elements of deep learning models - data, software, and hardware - we present examples of each component and describe how each has become accessible to internet users. By analyzing a set of deep learning papers published from 1987 to 2018, we find that deep learning research is becoming increasingly geographically decentralized: More institutions and countries around the world are publishing papers at prestigious deep learning conferences. We also find that more deep learning papers rely on a small set of publicly accessible datasets, such as thousands of facial images, and open-source software models for numerical computation and machine learning. This analysis of data, software, hardware, and geographic decentralization provides the background information necessary to establish how deep learning proliferation differs from that of other dual-use technology.

2.2 Background and history of deep learning

2.2.1 Subcategories of artificial intelligence

Artificial intelligence (AI) models are divided into categories based on their functionality and internal operation. In particular, machine learning is a subset of AI wherein data is used to improve a model's performance on a particular task.[9] Figure 2.1 provides more detail on various AI categories.[10]

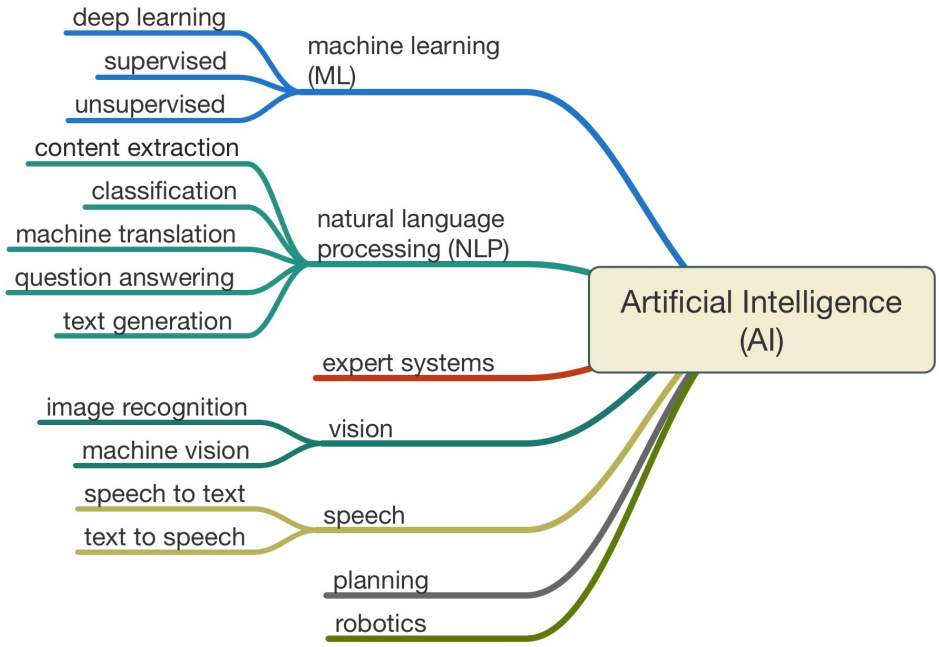

Figure 2.1: Various subcategories for artificial intelligence models.

As in Figure 2.1, deep learning is a subcategory of machine learning. In particular, deep learning relies on artificial neural networks to model data or make predictions. In the remainder of this section, we describe how artificial neural networks incorporate research in AI and neuroscience to model complex data, including sound, images, and videos.[11]

9. "What is Machine Learning? - Introduction | Coursera" 2019.

10. Gokani 2019.

11. Other categories in the model include natural language processing (NLP), expert systems, vision, speech, planning, and robotics. NLP systems focus on understanding, modelling, and generating text; expert systems are designed to reproduce the behavior of a human expert, such as a doctor or customer service agent. Although the vision category contains the specific applications of machine vision and image recognition, machine learning is often applied to these tasks as well. Similarly, while planning and robotics generally encompass algorithms for modelling real-world motion, such as moving robots or artificial limbs, machine learning algorithms may also be applied to this task.

2.2.2 History of deep learning and artificial neural networks

Deep learning models utilize artificial neural networks for training and making predictions. In the early 1940s, scientists studying the human brain began to write papers and build electrical models that simulate how neurons, the cells that transmit information in our brains, function.[12] Biological neurons connect a set of inputs, which encode information as electrical signals, to a cell body (known as the "soma"). The cell body processes input signals (called input activations) and releases output signals across synapses, which are the gaps between the output fibers of one neuron and the input fibers of another.[13] Artificial neurons reflect this biological blueprint by receiving a set of input activations, transforming them according to a set of weights, and outputting another signal.

A single neuron

The diagram below depicts the inputs, weights, and output of a single artificial neuron, which computes a mathematical transformation on a weighted combination of its inputs.

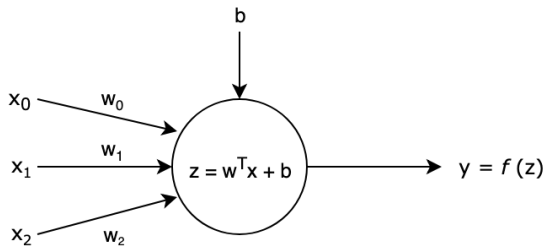

Figure 2.2: A single artificial neuron with three-dimensional input x, weights w, bias b, activation function $f(z)$, and output y.

The neuron portrayed above has three input signals x_0, x_1, and x_2, which are collectively denoted by the vector x. Each input x_i is associated with some scalar weight parameter w_i; this neuron has weights w_0 (associated with x_0), w_1, and w_2. The entire neuron also has an additional bias parameter b, which affects its final output y. The output y of the neuron is computed by taking the dot product of x and w, adding the bias term b, and using the result as the input to some nonlinear transformation function $f(z)$ (known as the activation function). Alternatively, y can be computed as $y = f(w^T x + b) = f(w_0 x_0 + w_1 x_1 + w_2 x_2 + b)$. Different weight vectors

12. "History of Neural Networks" 2019.
13. "Overview of Neuron Structure and Function - Molecular Cell Biology - NCBI Bookshelf" 2019.

w, bias terms b, and activation functions $f(z)$ allow neurons to compute a wide variety of output functions. Similar to a neuron in the human brain, the neuron receives a set of inputs, computes a mathematical transformation, and outputs a new signal.

Increasing research interest

In the 1950s, computer scientists began to program digital models of artificial neurons on early computers. However, as researchers frequently overstated but did not deliver on artificial neural networks' practical applications (some had heralded the imminent arrival of "artificial brains"), progress slowed. Some suggested that neural networks would be forever doomed by their computational complexity and inability to match performance of other learning models.[14]

In the 1980s, growing interest in biological and artificial neural networks prompted renewed research efforts; the first conferences on neural networks, including the Neural Information Processing Systems conference (NeurIPS, formerly NIPS),[15] were held in the late 1980s.[16] During this period, academic papers proposed new architectures for neural networks, which included assembling neurons into layers and using data points to optimize neurons' input weights.[17] These neural networks demonstrated promising results on machine learning tasks, including basic character recognition. Although artificial neural network research was subordinated to other promising machine learning models in the early 1990s, researchers in the mid-2000s rebranded the field as "deep learning" and published papers on the applications of neural networks with many layers - i.e. *deep* neural networks.[18]

Over the following decade, papers demonstrated how deep neural networks could match or outperform other machine learning models at a variety of tasks, from recognizing faces to detecting objects in images. As more efficient computing platforms were applied to deep learning, better results prompted research interest to increase, and new network architectures, software libraries, and computing hardware were developed to support these projects.[19]

14. "An Introduction to Neural Network Methods for Differential Equations | Neha Yadav | Springer" 2019.
15. In Section 2.3.3, we analyze all NeurIPS papers from 1987 to 2018 to track the number of universities, research labs, and countries performing deep learning research over the last twenty years.
16. "An Introduction to Neural Network Methods for Differential Equations | Neha Yadav | Springer" 2019.
17. "Perceptron Reading Material" 2019.
18. "A fast learning algorithm for deep belief nets. - PubMed - NCBI" 2019.
19. "Deep Learning 101 - Part 1: History and Background" 2019.

Assembling into networks

Artificial neural networks consist of interconnected layers of artificial neurons. Input data, such as images, text, or speech, is fed into neurons in the first layer. Each layer of artificial neurons receives input from the previous layer and computes successive nonlinear transformations on this input data before the network outputs a final value. Arranging single neurons into layers allows models to compute far more complex and expressive mathematical transformations. Below, we provide a diagram of a three-layer neural network with labels for relevant terms.

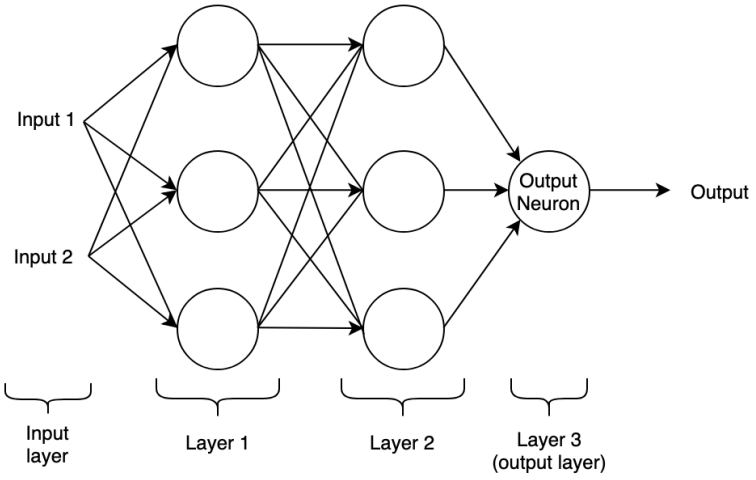

Figure 2.3: A basic three-layer neural network.

The neural network depicted in Figure 2.3 consists of an input layer and three layers of neurons. In this diagram, each input (there are two total) is connected to all neurons in the first layer. As a result, the neurons in Layer 1 will compute some transformation on the two inputs and subsequently transmit the result of this transformation to all neurons in Layer 2. Because all neurons in the first layer are connected to all in the second, the neurons in the second layer will compute a transformation on the outputs of the first layer and pass this output to Layer 3 - the output neuron.

To compute the output of a single neuron, the set of inputs are multiplied by their relative weights, summed, and subsequently used as the input to a mathematical transformation. The single neuron in Layer 3 computes the final output for the entire network. Compared to the neuron in Figure 2.2, this three layer network can compute more complex transformations across its three layers. More complex neural networks deviate slightly from this structure by performing more

sophisticated transformations on input data; for example, one layer may normalize inputs along a Gaussian distribution. Networks for object or facial recognition are more complex than the network in Figure 2.3 and may include millions of artificial neurons.

Types of deep learning models

Neural networks have proven particularly capable at processing high-dimensional data, such as images, text, and speech.[20][21] Given these capabilities, deep learning research has yielded specialized models for object recognition (determining what objects are in an image), facial recognition (matching unknown faces to known identities), speech recognition, and myriad other computer vision or artificial intelligence tasks.[22] Other deep learning research has led to innovative *generative* models, or algorithms capable of generating data that appears similar to a set of training data.[23] Based on the type of output produced by a particular model, we divide deep learning models into three categories:

1. **Classification**: Classification models output a discrete value that assigns an input variable to one of a set of known categories, where each category is known as a "class." Facial recognition is an example of a classification task: Input data, such as an image of an individual's face, is matched to a particular identity, or output class. Object recognition is another classification task: Given an image of some object, a model may assign it to one of many categories. Classification models frequently include an "unknown" class for inputs that do not match the set of known output classes.

2. **Regression**: Regression models produce a continuous output.[24] Linear regression - a statistical method for modeling a linear relationship among data points - represents a simple case. Other regression models may estimate the price of a home or evaluate the probability of a particular event occurring.

20. High-dimensional data refers to the amount of information contained in each data point. While some statistics - such as yearly population - may contain only one number in each sample, other data - such as images - require thousands or millions of numbers to represent only one piece of sample data. As a result, an image represents high-dimensional data and yearly statistics low-dimensional data.
21. "What Killed the Curse of Dimensionality? - Hacker Noon" 2019.
22. "The major advancements in Deep Learning in 2018 | Tryolabs Blog" 2019.
23. "Generative Models" 2019.
24. A continuous variable is not constrained to a set of possible outputs.

3. **Generative**: Generative models produce outputs similar to their training data. Unlike regression and classification models, which predict some output variable based on input data, generative models are designed to create data. For example, a generative model trained on facial images may learn to generate images that appear to be real human faces. Other generative models can produce art or music.[25] In Section 2.2.3, we describe how one category of generative models known as generative adversarial networks may be used to create deep fakes.

A neural network's complexity varies widely depending on its application; for example, Google's first Inception network, which provided high accuracy results for object recognition, utilized twenty-two layers and almost seven million parameters.[26] A more recent model created by OpenAI to generate text uses a over one billion parameters and a more complex neural architecture based on recurrent neural networks.[27] Section 2.2.2 provides greater technical detail on how certain types of neural networks are applied to facial recognition and creating deep fakes.

2.2.3 Models for facial recognition and deep fakes

Although more complex neural networks follow the same paradigm of repeated interconnected layers presented in Figure 2.3, specific neural network architectures are used for certain tasks. This section briefly provides greater technical detail on the deep learning models used to generate deep fakes and perform facial recognition, which are discussed in Chapters 4 and 5.

Convolutional neural networks and object recognition

Convolutional neural networks (CNNs), which are used for a variety of computer vision tasks, utilize specific mathematical functions and internal connections for transforming input data. Within a CNN, a convolutional layer of neurons computes weighted averages of subsets of the inputs; for example, when an image is used as input data, the first layer of a CNN may compute the average color value in various regions of the image.[28] Other layers in a CNN compute minima or maxima of subsets of the input data.

25. "[1812.04948] A Style-Based Generator Architecture for Generative Adversarial Networks" 2019.
26. "Going Deeper With Convolutions" 2019.
27. "Better Language Models and Their Implications" 2019.
28. "CS231n Convolutional Neural Networks for Visual Recognition" 2019.

Although CNNs were first proposed in the late 1990s, the 2012 NeurIPS paper "Imagenet classification with deep convolutional neural networks" (colloquially named "AlexNet" after its first author) demonstrated how a specifically architected CNN could yield high-accuracy results classifying the dataset ImageNet.[29] The original AlexNet paper has been cited over 36,000 times since its publication, and CNNs have been widely applied to myriad tasks from facial recognition to medical imagery analysis.[30]

Generative adversarial networks and deep fakes

Generative adversarial networks (GANs) harness competition between *two* neural networks - a generator network and a discriminator network - to produce data that appears similar to training samples. The discriminator network, which is able to determine whether a given sample input is "genuine" or "not genuine," is trained to recognize a particular class of data, such as real da Vinci paintings from fake ones. The generator, which is a generative model, uses feedback from the discriminator to progressively output samples that make the discriminator increasingly unsure whether a given example is real or fake. Once the GAN is trained, the generator produces output that makes the discriminator network unable to tell between genuine sample data and the generator network's output.[31] Given their ability to replicate input data, GANs have been applied to generating human faces that appear real,[32] creating artwork that appears in the style of famous artists,[33] and producing realistic deep fake videos.[34]

One paper titled "Recycle-GAN: Unsupervised Video Retargeting" uses generative adversarial networks to transfer speech and facial expressions from one video for another. Although this paper, which was published by researchers at Carnegie Mellon University in 2018, showcased the technology by transforming John Oliver's facial expressions into Stephen Colbert, it could be repurposed to generating deep fakes.[35] Alongside convolutional neural networks, GANs have been applied to tasks with clear national security implications.

29. "ImageNet Classification with Deep Convolutional Neural Networks" 2019.
30. "Alex Krizhevsky - Google Scholar Citations" 2019.
31. "A Beginner's Guide to Generative Adversarial Networks (GANs) | Skymind" 2019.
32. Bernal 2018.
33. Vincent 2019a.
34. "shaoanlu/faceswap-GAN: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping." 2019.
35. Bansal et al. 2018.

2.2.4 Training a neural network

Neural networks are trained using a dataset of training samples, where each sample includes a data point x_i with correct label $\hat{y}_i$. Each training sample x_i is fed into the network to generate a predicted output y_i during the training process. Then, given the network's current predicted output y_i and the correct output $\hat{y}_i$, the weights for every neuron in the network are adjusted to minimize the quantity $||y_i - \hat{y}_i||$, or the difference between the correct and predicted ouptuts for the sample x_i.[36] The process of feeding the example x_i through the network to generate a prediction during the training process is known as forward propagation; the process of adjusting neurons' weights to minimize $||y_i - \hat{y}_i||$ is known as backward propagation. Neural networks are frequently trained using the optimization procedure gradient descent, which performs forward and backward propagation repeatedly to minimize the model's error in correctly classifying training samples. Appendix 2.A discuses the specific mathematics required to train a neural network using forward ad backward propagation.

2.2.5 Explainable deep learning models

Although researchers have developed sophisticated methods for training high accuracy deep learning models, understanding neural networks' behavior remains challenging. Explainable models allow users to clearly interpret why an input sample generated a given output. However, given neural networks' high degree of interconnectedness and structural complexity, it is often difficult to explain how an input image with millions of pixels prompted a deep learning model to identify a given object or face. In a deep neural network, tens of layers may perform different transformations on distinct inputs, thereby making the entire model difficult to mathematically analyze. During the training process, each layer of the neural network generally learns to recognize increasingly complex features from the original input data.[37] For example, while the early layers of a facial recognition network may recognize basic lines and curves, subsequent layers may synthesize the output of previous layers to detect the presence or position of an individual's lips, eyes, and nose.

In March 2019, Google and OpenAI announced a collaborative effort to create "activation

36. Different metrics, including the L1 and L2 norm, may be used to measure this distance between predicted and correct output.

37. T. B. Lee 2018.

atlases," or pictorial representations of neural networks' intermediary layers designed to provide insight into models' "internal decision-making processes."[38] However, creating explainable deep learning models remains challenging for a variety of machine learning models.[39] In the next section, we discuss how data, hardware, and software are collectively harnessed to design and train deep learning models.

2.3 Data, software, and hardware

Deep learning systems require three elements: Data, hardware, and software. In recent years, all three components have all become more accessible to anonymous internet users. In the following sections, we provide an overview of how each component is used to create deep learning models. We subsequently analyze how publicly accessible datasets, open-source software libraries, and purpose built hardware have contributed to deep learning proliferation. We find that a growing percentage of deep learning research papers rely on publicly accessible datasets and open-source libraries, thereby posing challenges for regulating or controlling the technology.

2.3.1 Data

Two decades ago, machine learning datasets with millions of examples were virtually nonexistent.[40] Today, large datasets have become widely accessible on the internet, allowing virtually anyone to train sophisticated deep learning models. The amount of data required to train deep learning models varies widely depending on the application. Datasets for image recognition may include millions of samples; for example, ImageNet - one of the most widely used and cited image recognition datasets - has over fourteen million images with hundreds or thousands of images for each distinct class, including "vacuum cleaner" and "cappuccino."[41] Datasets for text-based deep learning systems may include hundreds of millions of words;[42] datasets for speech models may have hundreds of hours of audio recordings and transcripts.[43] Websites including data.gov and kaggle.com allow individuals

38. "Introducing Activation Atlases" 2019.
39. Ferris 2018.
40. "CS231n Convolutional Neural Networks for Visual Recognition" 2019; "ImageNet: the data that spawned the current AI boom - Quartz" 2019.
41. "ImageNet" 2019.
42. "1 Billion Word Language Model Benchmark" 2019.
43. "openslr.org" 2019.

to download datasets for machine learning, such as thousands of Chicago crime reports or millions of movie ratings.[44] Kaggle also hosts competitions with millions of dollars in prize money where individuals compete to submit the highest accuracy model. For example, the hedge fund Two Sigma hosted a competition with a $100,000 prize for using news data to predict stock price movements.[45]

Certain publicly available datasets have become industry-wide benchmarks for new deep learning models. When publishing papers on unconventional neural net architectures, researchers may report accuracy on these industry standard datasets in order to showcase their model's performance and robustness relative to previously published papers. For example, for image or object recognition models, researchers often report accuracy on ImageNet to assess how new systems compare to past models, such as Google's Inception model for object recognition.[46] Beyond concentrating research on certain datasets, these competitions have also led to increased enthusiasm from researchers and univerisities for sponsoring or collecting additional datasets. In Table 2.1, we outline the creator, purpose, and relative size of datasets frequently used in deep learning research.

In order to measure how open-source datasets have facilitated research and proliferation, we analyzed a dataset of 1,526 academic papers from 2013 to 2018 that have been published at one of a small group of selective deep learning conferences; these papers are included in the GitHub repository "Papers with code (PWC)."[47] Although it is relatively straightforward to find the code and models associated with a single academic paper, collecting a large dataset of 1,500 open-source papers was impractical. The "Papers with Code" repository links to 1,526 recent academic papers and their associated code repositories. All papers in the dataset were published at one of five academic conferences: The Conference on Computer Vision and Pattern Recognition (CVPR), the Conference on Neural Information Processing Systems (NeurIPS), the International Conference on Computer Vision (ICCV), European Conference on Computer Vision (ECCV), or International Conference on Machine Learning (ICML). One 2013 paper was published in the journal "Software - Practice and Experience," a peer-reviewed journal that focuses on software systems and applications.[48] Given these conferences' selectivity and popularity, the PWC dataset provided an effective starting point for collecting data on deep learning research.

44. "Datasets | Kaggle" 2019; "Data.gov" 2019.
45. "Competitions | Kaggle" 2019.
46. "[1602.07261] Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning" 2019.
47. "zziz/pwc: Papers with code. Sorted by stars. Updated weekly." 2019.
48. "Software Practice and Experience | RG Impact Rankings 2018 and 2019" 2019.

To analyze dataset usage in deep learning research, we wrote a Python program to download the code associated with these 1,526 published papers. Then, using another Python program, we searched through the code associated with each paper for references to widely used machine learning datasets, such as CIFAR and ImageNet. All datasets in Table 2.1 are publicly accessible (they can be downloaded on the internet for free). Academic websites, published papers, news articles, and open-source repositories were used to compile the data in Table 2.1.[49] In Appendix 2.C, we provide significantly greater detail on how the datasets and libraries listed in Tables 2.1 and 2.2 were selected.

Name	Content	Creator
ImageNet	Over 14 million images organized by WordNet syntactic hierarchy for object recognition	Professor Fei-Fei Li in 2007 at Princeton University
CIFAR (Canadian Institute For Advanced Research)	50,000 or 60,000 (depending on the version of the dataset) images for object recognition	Researchers at the University of Toronto
MNIST	60,000 training images of handwritten digits from 0-9	Collaboration among researchers at NYU, Microsoft, and Google
COCO (Common Objects in Context)	Millions of labels describing over 300,000 images; used for object detection and image segmentation	Collaboration among industry and academic researchers
SQuAD (Stanford Question Answering Dataset)	150,000 question and answer pairs created from content on Wikipedia	Researchers at Stanford University
HUB5	Transcripts of 40 recorded English conversations	Linguistic Data Consortium at the University of Pennsylvania
LSUN (Large-scale Scene Understanding)	Approximately ten million images of images with scene descriptors ("bedroom," "restaurant," "kitchen," etc.)	Princeton University

Table 2.1: Name, description, and creator for all datasets tracked.

49. "ImageNet" 2019; "CIFAR-10 and CIFAR-100 datasets" 2019; "MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges" 2019; Lin et al. 2014; "The Stanford Question Answering Dataset" 2019; "2000 HUB5 English Evaluation Transcripts - Linguistic Data Consortium" 2019; "LSUN" 2019.

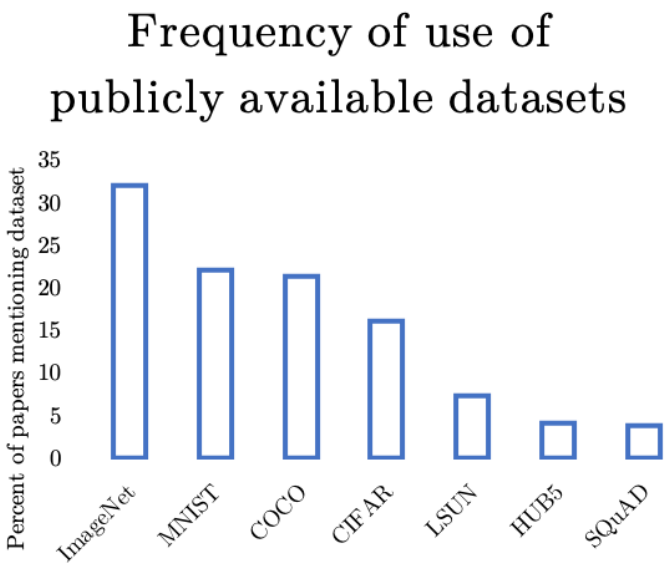

Figure 2.4: Dataset usage for all 1,526 papers from 2013 to 2018.

Figure 2.4 reports the results of tracking references to the datasets in Table 2.1 across the open-source models released by 1,526 recent academic papers from 2013 to 2018. The results in Figure 2.4 serve as a powerful demonstration of how a small set of nine publicly available datasets has become industry-standard among deep learning researchers. In all 1,526 papers analyzed, over *thirty percent* contained references to ImageNet, the fourteen-million image dataset often used to train object recognition models. Over twenty percent of papers referenced the MNIST (character recognition) and COCO (object and context recognition) datasets, another striking illustration of these datasets' critical role in academic research.

Why are these datasets so popular?

The accessibility, scale, and standardization enabled by the datasets in Table 2.1 provide critical context and justification for these results. Before ImageNet was released, researchers relied on a variety of smaller datasets that lacked a consistent structure for describing image content;[50] for example, a dataset of a few thousand images released by the FBI was often used for facial recognition by researchers in the early 2000s.[51] After ImageNet was first released in 2009, researchers began to vie for the highest accuracy model in annual competitions, and an increasing number of papers

50. "ImageNet: the data that spawned the current AI boom - Quartz" 2019.
51. "Face Recognition - FBI" 2019.

reported accuracy levels on the dataset. ImageNet continues to be an industry-standard benchmark for quickly evaluating model performance; for example, when Google published the fourth version of its image recognition model "Inception" in 2016, the paper's abstract reported 3.08% top-five error on ImageNet.[52][53]

These datasets' scale provides another explanation for their widespread use. Creating ImageNet required researchers to pay individuals for two years to collect and label images;[54] similarly, the researchers assembling the SQuAD dataset described in Figure 2.1 hired workers to produce over 100,000 question-answer pairs.[55] As more researchers adopted these datasets, tutorials and tools were created by companies and individuals to further increase ease of use. For example, Google Cloud Platform offers a step-by-step guide for training a deep neural network on the ImageNet dataset,[56] and Microsoft's Cognitive Toolkit software library provides a tutorial for training an image recognition model on the CIFAR dataset.[57] Beyond official company tutorials from Microsoft and Google, engineers and amateur users have created additional resources on personal websites or open-source code repositories.[58] Thus, the ease of access, number of samples, and emergence of industry-wide accuracy metrics precipitated the widespread usage of the datasets in Table 2.1.

2.3.2 Software

Researchers also rely on a small set of open-source software libraries for creating deep learning models. Software libraries contain a set of functions written in code for a particular purpose. For example, a "numerical computation library" may offer programmers access to a variety of prebuilt and pretested math functions. These libraries significantly simplify the process of designing and training neural networks. For example, using the software package Keras, it is possible to create simple neural networks in only a few lines of code. Appendix 2.B provides a short sample program for creating and training a neural network. In the remainder of this section, we discuss how open-source software libraries have further reduced technical barriers to training deep learning models.

52. Top five error indicates the percentage of examples where the model's five most likely outputs did not contain the correct result.
53. "[1602.07261] Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning" 2019.
54. "ImageNet: the data that spawned the current AI boom - Quartz" 2019.
55. "[1606.05250] SQuAD: 100,000+ Questions for Machine Comprehension of Text" 2019.
56. "Training ResNet on Cloud TPU | Cloud TPU | Google Cloud" 2019.
57. "Hands-on Labs Image Recognition - Cognitive Toolkit - CNTK | Microsoft Docs" 2019.
58. "How to get Images from ImageNet with Python in Google Colaboratory" 2019.

We focus specifically on the eight open-source deep learning libraries listed in Table 2.2 and track their use among the same dataset of 1,526 academic papers discussed in Section 2.3.1. Similar to our conclusions in the previous subsection about research practices for deep learning datasets, software library usage clearly reflects a progression towards reliance on a small set of open-source repositories. In indicating a clear trend towards using open-source libraries, these results definitively indicate how free and accessible software facilitates cutting-edge deep learning research.

Understanding how researchers use open-source software presents clear insights into tracing how deep learning technology is created and spreads. As with our experiments on datasets, we wrote a Python program to search for machine learning library usage in the code associated with the 1,526 recent machine learning academic papers published at a small set of selective machine learning conferences that was previously described in Section 2.3.1. The research papers used a variety of coding languages; however, in collecting data on open-source libraries, we searched only Python files for library usage. Python has become very popular for machine learning and data science applications and is easily searchable for references to libraries or datasets.[59] In Table 2.2, we provide an overview of the libraries tracked in these experiments. News articles, blog posts, open-source code repositories, and deep learning tutorials were used to compile the data in Table 2.2.[60] Appendix 2.C provides significantly greater detail on why we chose to focus on the software libraries listed in Table 2.2. In addition to those in the table, Chinese technology company Baidu open-sourced its deep learning software PaddlePaddle; however, in the set of papers we analyzed from the "Papers with code" repository, none appeared to use PaddlePaddle so it was not included in this table.[61]

59. Patel 2018.

60. "Announcing PyTorch 1.0 for both research and production" 2019; "TensorFlow – opensource.google.com" 2019; "BVLC/caffe: Caffe: a fast open framework for deep learning." 2019; "Caffe | Deep Learning Framework" 2019; "The Microsoft Cognitive Toolkit - Cognitive Toolkit - CNTK | Microsoft Docs" 2019; "MXNet: A Scalable Deep Learning Framework" 2019; "Introduction to the Python Deep Learning Library Theano" 2019; "deepmind/sonnet: TensorFlow-based neural network library" 2019.

61. Vincent 2016.

Name	Purpose	Creator and maintainer	Dependencies
PyTorch	Python library for performing matrix operations or modeling deep neural networks	Facebook AI	Standalone machine learning library
Tensorflow	Provides a variety of data processing and numerical computation functions for machine learning	Google Brain	Standalone numerical computation library
Caffe	Modular deep learning framework particularly optimized for speed; models are easy to share and are often used for computer vision tasks	Created by UC Berkeley PhD student Yangqing Jia; maintained by Berkeley AI group	Standalone machine learning library
Keras	Neural network library written in Python; can be used with Tensorflow, Microsoft Cognitive Toolkit, or Theano	Created by Francois Chollet, an engineer at Google	Requires Tensorflow, Microsoft Cognitive Toolkit, or Theano
Microsoft Cognitive Toolkit	Provides a deep learning library capable of scaling across multiple machines or programming languages	Microsoft Research	Standalone machine learning library
MXNet	Scalable deep learning framework that can perform efficient task scheduling and support multiple programming languages	The Apache Software Foundation	Standalone machine learning library
Theano	Similar core functionality to Tensorflow; provides a framework for defining and evaluating matrix expressions	Developed by the Montreal Institute for Learning Algorithms at the University of Montreal, Quebec	Standalone numerical computation library
Sonnet	Creates hierarchy of models and neural architectures to allow for easy experimentation	DeepMind, Alphabet's deep-learning subsidiary	Requires Tensorflow

Table 2.2: Widely used machine learning libraries tracked in our analysis of open-source academic software. It is important to note that while organizations - such as Facebook AI or DeepMind - may maintain particular libraries - such as PyTorch and Sonnet - anyone with internet access may contribute to open-source software development.

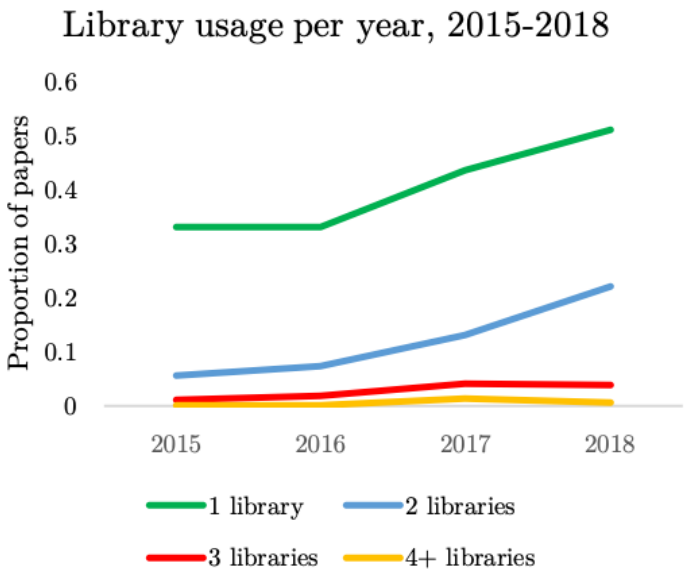

Figure 2.5: Library usage per year.

Figure 2.5 shows how researchers have increasingly utilized the open-source libraries in Table 2.2 in published research. For example, the paper "An Improved Deep Learning Architecture for Person Re-Identification," which was presented at the 2015 Conference on Computer Vision and Pattern Recognition (CVPR), uses Tensorflow and Keras to create a neural network specifically designed to re-identify individuals in sequences of images. Another paper titled "SSH: Single Stage Headless Face Detector," which was published in the 2017 IEEE International Conference on Computer Vision, used the library Caffe to implement an innovative structure for a neural network capable of identifying faces. Using open-source software for deep learning dramatically reduces the work required to develop cutting-edge models: Instead of designing and testing their own neural network or numerical computation libraries, research teams and companies may simply rely on a common codebase of trusted tools. Of the 1,526 papers analyzed, 1,159 contained Python scripts. 527 papers (of all 1,526) used none of the open-source libraries presented in Table 2.2; all 999 remaining papers (65.4% of the papers in "Papers with code") used at least one open-source software library in Table 2.2.

Recent updates to software libraries have facilitated the upward trends in Figure 2.5. Microsoft's Cognitive Toolkit was first released in 2016; since then, the library has been updated to include more

efficient convolutional layers and compatibility with other mathematical computation libraries.[62] Microsoft also has released significant documentation on using Cognitive Toolkit with its Azure cloud computing platform, thus reducing barriers to entry and the costs associated with training deep learning systems.[63] Similarly, Google offers lengthy documentation on how to use Tensorflow on Google Cloud Platform (GCP); tutorial topics include how to train an image recognition network and how to run Google's Tensorflow library on distributed systems.[64] Each article provides a step-by-step outline for setting up a GCP machine for machine learning, includes an estimated cost for running the tutorial, and clearly walks users through relevant commands and lines of code.[65] Because these tutorials make it easier to conduct deep learning research, they further explain why Figure 2.5 suggests that software libraries have become more popular.

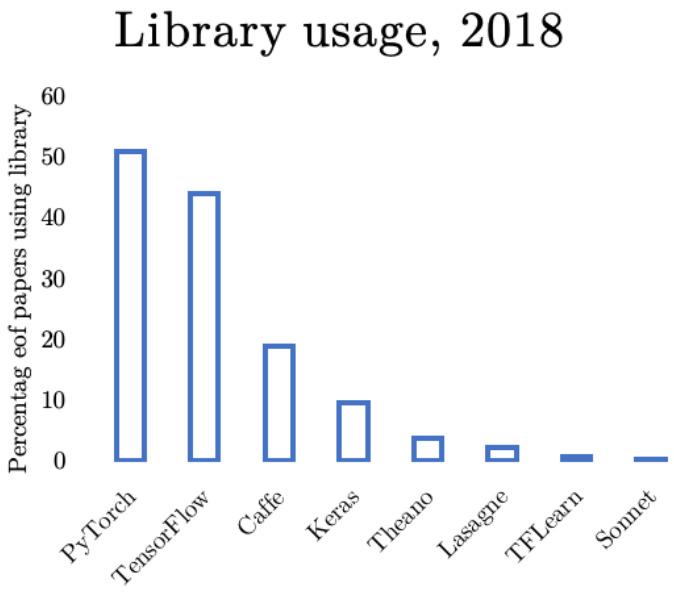

Figure 2.6: Libraries used in 2018.

Figure 2.6 provides a breakdown of library usage by all papers in 2018. Similar to academic papers' reliance on a small set of datasets (such as over 30% of papers mentioning ImageNet), two deep learning libraries (PyTorch and Tensorflow) are used in over 40% of the open-source code published by papers in 2018. While slightly over 50% of papers used PyTorch and over 40%

62. "CNTK_2_6_Release_Notes - Cognitive Toolkit - CNTK | Microsoft Docs" 2019.
63. "CNTK on Azure - Cognitive Toolkit - CNTK | Microsoft Docs" 2019.
64. "Deep Learning VM Image | Deep Learning VM Image | Google Cloud" 2019.
65. "Image Classification using Flowers dataset | Cloud ML Engine for TensorFlow | Google Cloud" 2019.

used Tensorflow, under 10% used Keras, Lasagne, and Theano. Slightly under 20% used the deep learning framework Caffe. Given Tensorflow's age (it was open-sourced in late 2015[66]) and level of integration with Google Cloud Platform, its popularity is unsurprising. The frequency of references to PyTorch is also unsurprising. The software library "Torch" was originally released on 2002 to provide a variety of functions for scientific computation, matrix math, and machine learning;[67] in 2016, researchers at Facebook released PyTorch, an open-source Python implementation of Torch that provided similar functionality to the original library.[68] Other libraries in Figure 2.6 are not nearly as popular as PyTorch and TensorFlow. Each library's unique use case and age may explain these results. For example, although the library Sonnet was created by Alphabet's subsidiary DeepMind, it was released more recently than any other library listed in Table 2.2. Appendix 2.C provides more detail about why we selected the datasets in Table 2.1.

As revealed by our analysis of dataset usage, a small set of software libraries are used extensively by deep learning research groups. Once these libraries are released and downloaded, it is impossible to control their use. Furthermore, thousands of developers may contribute to open-source libraries: Tensorflow has been modified by over 1,900 individuals[69] while PyTorch has been modified by over 1,000.[70] As the research and open-source communities utilize and develop the same set of software tools, it becomes more challenging to track their use.

Relationships among software libraries

Dependencies among software libraries provide another consideration for understanding the role software libraries play in deep learning proliferation. While some libraries listed in Table 2.2 were designed as standalone projects for supporting machine learning research, others were developed to work in conjunction with other open-source software. The rightmost column of Table 2.2 provides a brief overview of these dependencies. For example, Sonnet, which is developed by Alphabet's subsidiary DeepMind, requires a user to also install Tensorflow.[71] Other libraries possess similar dependencies; for example, Keras, another neural network library, requires that Tensorflow, Theano,

66. Metz 2015.
67. Scarpino 2018.
68. "PyTorch Releases Major Update, Now Officially Supports Windows" 2019.
69. "tensorflow/tensorflow: An Open Source Machine Learning Framework for Everyone" 2019.
70. "pytorch/pytorch: Tensors and Dynamic neural networks in Python with strong GPU acceleration" 2019.
71. "Open sourcing Sonnet - a new library for constructing neural networks | DeepMind" 2019.

or Microsoft Cognitive Toolkit also be installed.[72] These dependencies suggest that future libraries may follow a similar model of building on existing open-source tools.

2.3.3 Hardware

Multiple technology companies have announced or released hardware products specifically designed for deep learning. As these systems reduce the amount of time or electricity required to train and use neural networks, hardware represents another critical element of deep learning proliferation. Deep learning hardware has become more widely available on cloud platforms and through direct sales.

What is deep learning hardware?

Deep learning hardware is designed to accelerate numerical computations, such as multiplying matrices. As neural networks may have tens of thousands of distinct inputs and millions (or billions) of weight parameters, components of the training and testing process can be succinctly represented by matrix operations. Because individual matrix operations (such as multiplication) can be performed in parallel, computing hardware capable of performing thousands of parallel operations is well-suited to deep learning.

Graphics processing units (GPUs), which can process hundreds or thousands of mathematical operations at once, are very popular for training deep learning models. While the central processing units (CPUs) in desktop and laptop computers may be able to process two or four instructions in parallel, the latest GPUs may have hundreds or thousands of distinct hardware units capable of simultaneously executing mathematical instructions.[73] However, as the CPU is responsible for running critical elements of the operating system, CPUs are generally able to execute more complex instructions for storing and processing data. In contrast, GPUs are generally designed to carry out a larger number of simple instructions in parallel.[74] As matrix math often involves repeated addition and multiplication - which represent fairly simple mathematical operations - GPUs can provide much greater parallelization and training speeds.[75] By increasing training speeds, better hardware

72. “Home - Keras Documentation” 2019.
73. “How many cores in a standard gpu? - Quora” 2019.
74. “What is the difference between CUDA core and CPU core? - Stack Overflow” 2019.
75. “GPU computing: Accelerating the deep learning curve | ZDNet” 2019.

makes deep learning cheaper, less time intensive, and more accessible.

Recognizing that hardware and software architectures can be optimized to expedite neural network training, companies have begun to release hardware and software products specifically designed for deep learning. For example, NVIDIA's CUDA parallel computing platform allows programs to specify operations that should be run in parallel on separate GPU cores.[76] In recent years, NVIDIA, Amazon, Google, Intel, and smaller technology startups have designed or proposed hardware designed specifically for accelerating machine learning. These hardware platforms offer orders of magnitude higher performance and energy efficiency; for example, Google's Tensor Processing Unit (TPU) chips provide 15-30 times faster performance and 30-80 times higher performance per-watt than existing CPU and GPU hardware.[77] In Table 2.3, we compare hardware specifically designed for deep learning; news articles, company announcements, and blog posts were used to compile this data.[78]

New hardware's role in proliferation

While the new hardware listed in Table 2.3 may promise multiple orders of magnitude acceleration for deep learning computations, *any computer* is capable of training deep learning models, albeit at a much slower rate. Some studies have examined the relative speeds of training models on laptop and desktop computers; while laptops with graphics processing units are often used to train models for research and commercial applications, desktop computers and cloud computing platforms may provide one or more orders of magnitude faster training speeds.[79] One study found that using a single NVidia 1080 Ti GPU (sold for $699 in March 2019[80]) yielded training speeds 167 times faster than the Intel i5 CPU on a Macbook Pro.[81] When performing the experiments described in Chapter 5, we trained deep learning models on a laptop computer and on servers via Google Cloud Platform.

Multiple technology companies offer cloud computing platforms optimized for performing deep

76. "CUDA Zone | NVIDIA Developer" 2019.
77. "An in-depth look at Google's first Tensor Processing Unit (TPU) | Google Cloud Blog" 2019.
78. "An in-depth look at Google's first Tensor Processing Unit (TPU) | Google Cloud Blog" 2019; "Announcing AWS Inferentia: Machine Learning Inference Chip" 2019; "AWS announces new Inferentia machine learning chip | TechCrunch" 2019.
79. "TensorFlow performance test: CPU VS GPU - Andriy Lazorenko - Medium" 2019.
80. "GeForce GTX 1080 Ti Graphics Cards | NVIDIA GeForce" 2019.
81. "Benchmarking Tensorflow Performance and Cost Across Different GPU Options" 2019.

Hardware name and creator	Description	Available for purchase?	Available for cloud computing?	Performance improvement
Google Tensor Processing Unit (TPU)	Hardware chips designed to accelerate machine learning; specifically for Google's TensorFlow library	No	Can be rented on Google Cloud Platform since June 2018	Stated 15-30 times faster than existing CPU and GPU technology
Amazon Inferentia	Hardware chip capable of accelerating computations for multiple machine learning libraries	Unknown	Release planned on Amazon Web Services in 2019	Unknown
Unnamed chip created by Intel and Facebook	Unknown	Unknown; release planned for late 2019	Unknown; release planned for late 2019	Unknown
NVidia Tesla V100 GPU accelerator	Graphics processing chip designed to accelerate machine learning	Yes; each GPU sold for $6,099.00	Yes via Amazon Web Services and Google Cloud Platform	Similar performance to Google TPUs; less energy efficient
NVidia DGX-2	350 lb server containing 16 NVidia Tesla V100 GPUs; designed and marketed for deep learning	Yes; each unit priced at $99,000	N/A	Stated performance "equivalent of 300 servers with dual Intel Xeon Gold CPUs costing over $2.7 million dollars"

Table 2.3: Hardware designed to accelerate machine learning.

learning. Microsoft, Amazon, IBM, Oracle, Google, and Alibaba all offer computing resources through online cloud platforms; while Amazon Web Services (AWS) and Microsoft Azure lead in market share (AWS and Azure market shares are estimated to be above 30% and 15% market respectively[82]), Microsoft, Google and Alibaba have reported high growth rates. Over a one year period from 2017 to 2018, Microsoft reported 76% growth in Azure revenue[83] while AWS earnings rose 45% in Q4 2018.[84] Although research suggests that cloud computing costs are not dropping dramatically (one study found that over five years, Google Cloud Platform computing prices have dropped around 20% while other companies have maintained stable prices),[85] more cloud platforms and setup resources increase access and ease of use. Amazon, Microsoft, Google, IBM, Oracle, and Alibaba have published tutorials and documentation specifically for running deep learning models on their cloud platforms.[86] By attracting amateur programmers, corporate clients, and research institutions to cloud computing infrastructure, these tutorials may yield significant contributions to a technology company's bottom line. Thus, even though the price of cloud computing has not dropped significantly, increased software compatibility, better tutorials, and purpose-built machine learning hardware facilitate more deep learning research and more widespread use.

2.4 Geographic decentralization

Understanding deep learning's geographic decentralization provides additional information for comparing it to other dual-use technologies. In particular, as research and application become increasingly distributed around the world, national governments may struggle to mitigate harmful applications. In order to approximate geographic growth in deep learning research, we analyzed a dataset consisting of all papers from the Neural Information Processing Systems (NeurIPS) conference from 1987 to 2018. NeurIPS, which accepts roughly 20% of papers,[87] is widely considered to be one of the most prestigious and well attended conferences in artificial intelligence research; in

82. "AWS vs Azure vs Google Cloud Market Share 2018 Report" 2019.
83. "Microsoft Azure Revenues Climb 76% | Light Reading" 2019.
84. "AWS Cloud Revenue Jumps 45% in Q4, Microsoft Azure Revenue Up 76%" 2019.
85. "Cloud costs aren't actually dropping dramatically - Hacker Noon" 2019.
86. "Get Started with Deep Learning Using the AWS Deep Learning AMI | AWS Machine Learning Blog" 2019; "Deep Learning and AI frameworks - Azure | Microsoft Docs" 2019; "Deep Learning VM Image | Deep Learning VM Image | Google Cloud" 2019; "Getting started tutorial" 2019; "Get Started with Oracle Machine Learning" 2019; "Deep learning - User Guide| Alibaba Cloud Documentation Center" 2019.
87. "NIPS Accepted Papers Stats - Machine Learning in Practice" 2017.

recent years, leading institutions publishing papers at NeurIPS include Google, Carnegie Mellon, MIT, Microsoft, and Stanford. Given NeurIPS' longstanding relevance, popularity, and rigor, it proved a logical choice for measuring geographic decentralization. Furthermore, a user on GitHub released a code repository to track and download all NeurIPS papers from a certain year range.[88] After verifying that this code successfully downloaded the papers for a particular year, we modified the software to compile a dataset of all papers from 1987 to 2018. After using open-source software to extract the stated affiliations of each author, we were able to track the number of institutions contributing to published research at NeurIPS each year. See Appendix 2.C for more information how author affiliations were extracted and aggregated into countries and institutions.

The data clearly indicates that more organizations have published papers on deep learning, perhaps motivated by promising applications or the shrinking technical and monetary barriers to conducting research.

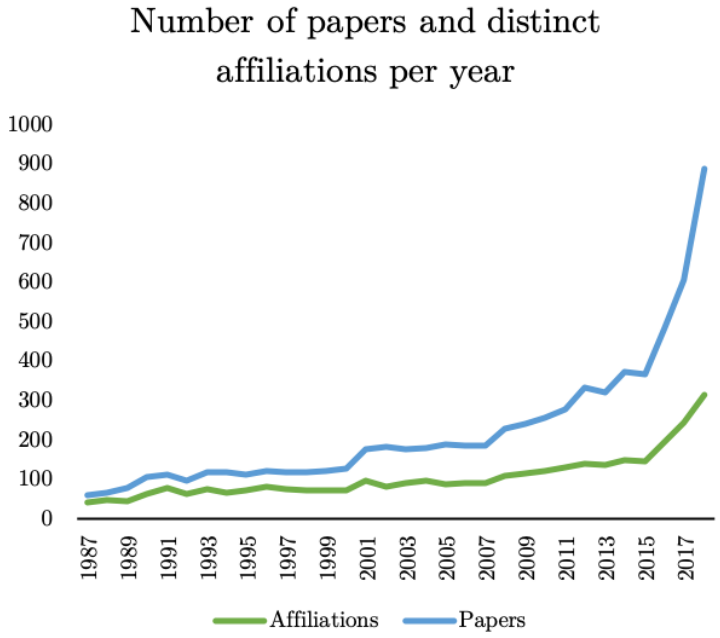

Figure 2.7: Number of papers submitted to NeurIPS per year and number of distinct author affiliations. Note that only papers where the author affiliation could be extracted are included in this chart.

Figures 2.7 demonstrates rapid growth in the number of distinct institutions conducting deep learning research. In addition to growth in the number of NeurIPS papers submitted by some

88. "benhamner/nips-papers" 2019.

non-US countries (such as China), growth in the number of research labs and companies in the United States is also high. From 2017 to 2018, the number of papers accepted by NeurIPS increased 46% from 606 to 887; while the number of papers published by the United States grew 38% from 299 to 413, publications by other countries, including China, and Israel, also grew notably.[89] Our program calculated that the number of papers from Chinese institutions accepted by NeurIPS grew from 8 to 19 from 2017 to 2018, and the number of papers published by Israeli institutions doubled from 5 to 10 in the same period. Because these numbers appeared remarkably small compared to US publication figures for NeurIPS, it appears possible that our program had a lower accuracy rate for identifying the author names and affiliations for non-US institutions, which could contain special or difficult to recognize characters. However, data collected by other researchers on NeurIPS publications reveals a similar imbalance between the US and rest of the world.[90]

This data poses clear implications for analyzing deep learning proliferation. While the development of some dual-use technology - such as nuclear weapons or rocket engines - can be geographically contained, deep learning research has now spread throughout tens of countries and hundreds of institutions. In Chapter 4, we study this proliferation process in the context of deep fakes, which were originally developed by academia but quickly transitioned into open-source and accessible software.

2.4.1 How are deep learning models shared?

In this subsection, we briefly describe how companies, universities, and individuals share or publish deep learning models on the internet. Software can be shared on the internet using a variety of free online collaborative software development platforms, such as GitHub, BitBucket, and GitLab. Individuals can also host models or code on a personal website. GitHub, BitBucket, and GitLab allow internet users to create a free account that can be used to upload code and documentation to public or private projects; public projects can be viewed, downloaded, or modified by other users.[91]

As these online platforms progressively track every addition and deletion to and from software libraries, they allow thousands of developers around the world to review each other's contributions. Many of the largest and most popular open-source projects, such as the website component library

89. Other researchers have used citation counts and publication counts to compare national progress in AI research. Although differing publication standards may affect the validity of these metrics, growth in the number of affiliations publishing papers - as in Figure 2.7 - may not be as affected.

90. "Who's Ahead in AI Research? Insights from NIPS, Most Prestigious AI Conference" 2019.

91. "GitHub vs. Bitbucket vs. GitLab vs. Coding - flow.ci - Medium" 2019.

Name	Approximate number of developers using service	Approximate number of organizations using service
GitHub	Over 31 million (2018)	Over 2.1 million organizations (2018)
BitBucket	5.0 million (2016)	"900,000 teams" (2016)
GitLab	No public data available	"More than 100,000 organizations" (2018)

Table 2.4: Different online platforms for sharing software.

Polymer and operating-system Linux, are hosted on GitHub and have been modified by thousands of individuals. Before a contribution is approved, other programmers or project managers will review new code, offer feedback, and scrutinize potential vulnerabilities. Selected services offered by the code collaboration platforms GitHub, BitBucket, and GitLab are contrasted in Table 2.4. Financial reports and company blog posts were used to compile the data in this table.[92]

GitHub, BitBucket, and GitLab are widely used by research labs, companies, and individuals to share models, libraries, or projects. For example, the Stanford Natural Language Processing and Machine Learning Groups maintain GitHub websites with open-source models, such as CoreNLP, which can compute the part of speech of words in a sentence;[93] Google maintains a GitHub website with almost 1,500 project repositories, from graphics rendering libraries for the Android operating system[94] to neural network models used in published research.[95] In 2018, over thirty-one million developers published code to GitHub, which now stores over ninety-six million projects.[96] Over ten thousand individuals have contributed to GitHub's most popular open-source projects, which include Microsoft's VSCode editing application and Facebook's React Native library for building mobile applications. Individuals may use GitHub to host personal projects,[97] contribute to open-source software, or privately share code for collaborative development among small teams. The scale of contributions to these platforms is striking: As technology proliferates through online

92. "The State of the Octoverse | The State of the Octoverse reflects on 2018 so far, teamwork across time zones, and 1.1 billion contributions." 2019; "Document" 2019; "Announcing $100 million in Series D round funding led by ICONIQ Capital | GitLab" 2019; "GitHub vs. Bitbucket vs. GitLab vs. Coding - flow.ci - Medium" 2019.

93. "Stanford NLP" 2019; "Stanford Machine Learning Group" 2019.

94. "google/filament: Filament is a real-time physically based rendering engine for Android, iOS, Windows, Linux, macOS and WASM/WebGL" 2019.

95. "google/uis-rnn: This is the library for the Unbounded Interleaved-State Recurrent Neural Network (UIS-RNN) algorithm, corresponding to the paper Fully Supervised Speaker Diarization." 2019.

96. "The State of the Octoverse | The State of the Octoverse reflects on 2018 so far, teamwork across time zones, and 1.1 billion contributions." 2019.

97. Personal projects include coursework, websites, or new open-source repositories.

platforms with millions of anonymous users, how can policymakers track or curb malicious use? This question reveals why deep learning poses challenges unlike other dual-use technologies

2.5 Norms in deep learning research

The technological components discussed in this chapter - data, software, and hardware - have contributed to norms of anonymity and open-sourcing in deep learning research and application. This section highlights how these two norms distinguish deep learning proliferation from other dual-use technology and thus pose unique challenges for policymakers.

Anonymity

GitHub, BitBucket, and GitLab, which are online platforms for sharing code, permit anonymous users to make contributions to open-source projects or download software. Many GitHub accounts, which allow individuals to publish or contribute to open-source projects, lack real names; false identifications or "throwaway" emails may also be used to create accounts on the platform. Because access to a GitHub account allows individuals to download software, anonymity makes it impossible to track the set of users downloading or contributing to open-source software. Chapter 4 further explores how anonymity was used by individuals developing and releasing deep fakes.

Open-sourcing

Open-sourcing characterizes deep learning software development in both industry and academia. From the operating system Linux to the internet browser Firefox, open-source projects allow contributors from around the world to develop a common project. Today, open-sourcing is common among deep learning research for spurring new breakthroughs and ensuring reproducibility. Other areas of computer science research also rely on open-sourcing; for example, security researchers often rely on open-source algorithms for encryption and decryption to ensure transparency and testing.[98] However, as anyone may download or experiment with open-source software, it is effectively impossible to control or track.

98. Goodin 2019.

2.5.1 Conclusion: Lowering the technical bar and increasing decentralization

Datasets, software libraries, and hardware have become increasingly distributed, accessible, and valuable for conducting deep learning research. As these developments have resulted in norms of anonymity and open-sourcing, policymakers will struggle to limit open-source software contributions or track the anonymous internet users publishing deep learning models. This analysis suggests that the "cat is out of the bag" with respect to creating deep learning systems. Thus, should policymakers focus on creating impediments or delays to prevent individuals from pursuing particular deep learning applications? The following chapters pursue a deeper analysis of deep learning proliferation for the purpose of answering this question: If malicious applications cannot be systematically prevented, can they be delayed or disrupted?

Appendices

2.A Neural networks and backpropagation

This appendix provides a cursory overview of the mathematics required to train a simple neural network; numerous online resources, from course material to YouTube videos, provide far more detail and resources on this subject.

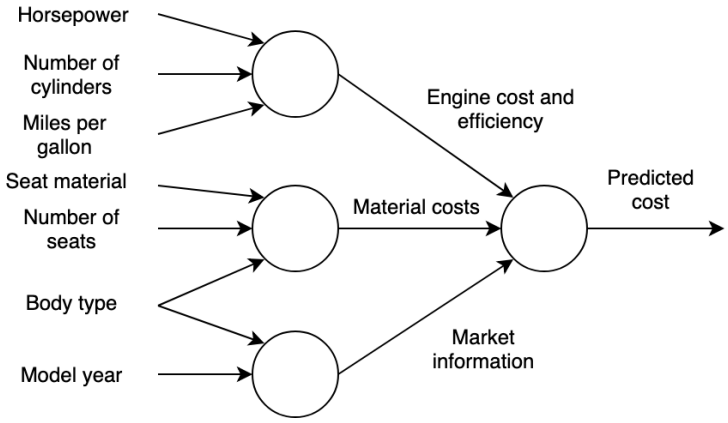

Figure 2.A.1: A two-layer neural network with seven inputs.

Figure 2.A.1 provides a diagram of a two layer neural network that could be used to predict the price of a car given information on the engine, interior, and model. In the first layer, the top neuron takes three inputs: The engine's horsepower, the number of cylinders in the engine, and the vehicle's mileage per gallon; the middle neuron has three different inputs - seat material, the number of seats, and the body type; and the bottom neuron in the layer is provided with the vehicle's body type and model year. As each neuron in this layer associates different weights with each input, the uppermost neuron is essentially taking measurements of the engine cost and vehicle efficiency, the center neuron is measuring the material and interior cost of the car, and the bottom

neuron assesses broader market factors, such as the price of different types of vehicle bodies (such as sedans, SUVs, and coupes). The single neuron in the second layer is provided with the outputs of the previous layer.

Training this network requires a dataset containing vehicle information (including vehicle prices) and choosing a loss function $\mathscr{L}$ that will measure the network's ability to accurately predict vehicle price. In this example, we use the square loss function, which is calculated using Equation (2.1) for a single training example. In this equation, $\hat{y}^{(i)}$ is the networks' predicted price on training example i and $y^{(i)}$ is the true vehicle price for training example i. $\mathscr{L}^{(i)}$ is the loss on training example i, or the distance between the desired output and the network's current output.

$$\mathscr{L}^{(i)} = (\hat{y}^{(i)} - y^{(i)})^2 \tag{2.1}$$

After initializing the weights and bias parameters in each neuron, training is performed in two steps: Forward and backward propagation. In the forward propagation step, each piece of training data, such as the information regarding a single car, is provided as input to the neural network, fed through each layer of neurons, and used to generate a final predicted cost $\hat{y}$. As the true vehicle price y is known for all training data points, the weight parameters for the entire neural network can be updated to minimize the loss $\mathscr{L}^{(i)}$, or the difference between y and $\hat{y}$. Below, we provide greater detail on updating the weights to train the output neuron.

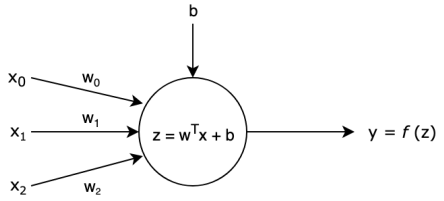

Figure 2.A.2: The output neuron of the vehicle price prediction neural network.

Updating the weights for this neuron entails computing the derivatives of the loss with respect to each weight. In Figure 2.A.2, we provide another diagram of a single neuron to show how these derivatives are computed and used to adjust weights in order to reduce the loss for a given training example. The four derivatives relevant to the neuron in Figure 2.A.2 are $\frac{\partial \mathscr{L}^{(i)}}{\partial w_0}$, $\frac{\partial \mathscr{L}^{(i)}}{\partial w_1}$, $\frac{d \mathscr{L}^{(i)}}{d w_2}$, $\frac{\partial \mathscr{L}^{(i)}}{\partial b}$. For simplicity, we will compute only the first term, or the gradient of the loss with respect to w_0.

We use the ReLU activation function, or $f(z) = \max(z, 0)$, Thus, the output neuron computes:

$$\hat{y} = f(z) = \max(w_0 x_0 + w_1 x_1 + w_2 x_2 + b, 0)$$

We now compute the gradient of the squared loss for one training example.[99]

$$\frac{\partial \mathscr{L}^{(i)}}{\partial w_0} = \frac{\partial \mathscr{L}^{(i)}}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial w_0} = 2\left(y^{(i)} - \hat{y}^{(i)}\right) x_0 \, 1\left\{w^T x + b > 0\right\}$$

Using gradient descent, we can update the weight w_0 by setting $w_0 = w_0 - \alpha \frac{\partial \mathscr{L}^{(i)}}{\partial w_0}$ where α is some adjustable training parameter. The other weights for this neuron can be updated similarly; the parameters for the three neurons in the first layer are updated using the same procedure but with additional derivative terms. The process outlined above is known as backward propagation as parameters are updated during a backwards progression through the network from the output neuron.

2.B Simple neural network in Keras

In this appendix, we discuss how a simple two-layer neural network could be implemented in software using the neural network library Keras. The first block of code below defines the mathematical structure and optimization settings for the network.

```
model = Sequential()
model.add(Dense(3, input_dim=7, activation='linear'))
model.add(Dense(1, activation='linear'))
model.compile(loss='mean_squared_error', optimizer=sgd, metrics=['accuracy'])
```

The first line defines the model as a sequential network of interconnected layers; the second and third line add layers of neurons to this model. The fourth line instructs Keras to use the mean square error metric for evaluating the model's performance on training data, and to use stochastic gradient descent - an optimization strategy for learning to fit data - to train the network.

99. We use the notation $1\left\{w^T x + b > 0\right\}$ as an indicator variable that is 1 when $w^T x + b > 0$ and 0 otherwise.

```
model.fit(X_train, Y_train, epochs=100, batch_size=10)
results = model.evaluate(X_test, Y_test)
```

The first line provides training examples in the matrices X_{train} and Y_{train}. While the X_{train} matrix stores the data associated with a training sample, Y_{train} contains the correct labels for each example in X_{train}. For example, X_{train} may store a set of images while Y_{train} contains a list of objects in each image. The additional two parameters - epochs and batch size - control how many examples are used in each training step and how long the model will be trained. At this point, running this program with actual training data will result in the model learning to fit X_{train} and Y_{train}.

2.C Dataset selection methodology

This appendix discusses how the datasets in Table 2.1, software libraries in Table 2.2, and machine learning hardware in Table 2.3 were selected.

Dataset selection methodology

Academic papers, news articles, and educational curricula served as the rationale for selecting the datasets in Table 2.1. Online deep learning tutorials, such as fast.ai,[100] characterize MNIST, ImageNet, and CIFAR as "household name" datasets for computer vision tasks. Because the field of deep learning is far broader than computer vision, we included other datasets as well, including some for automated question answering and speech recognition. Multiple articles and course materials referenced the SQuAD and HUB5 datasets in these areas of machine learning.[101]

Open-source library selection methodology

We selected the libraries in Table 2.2 after surveying academic papers, commercial software usage, and open-source contributions. Tensorflow, Caffe, PyTorch, Adobe MXNet, Microsoft Cognitive

100. "fast.ai Datasets | fast.ai course v3" 2019.
101. "Swift Documentation Quick Guide" 2019; "Fueling the Gold Rush: The Greatest Public Datasets for AI" 2019.

Toolkit, and Keras are almost universally cited by blogs and deep learning tutorials for their popularity and extensive functionality.[102]

Although the popular library Theano is used in some applications and academic papers, it is no longer actively developed;[103] we chose to include it in our dataset because other deep learning frameworks, such as Keras and Lasagne, are still compatible with Theano's computation functions.[104] Furthermore, while Sonnet, the deep learning library released in 2017 by Alphabet subsidiary DeepMind, is not nearly as popular as Keras or PyTorch, its recent release and ongoing support from leaders in deep learning research suggest that it will be used for both open-source and academic purposes.[105]

Hardware selection methodology

The set of hardware components outlined in Table 2.3 was compiled from various media articles, conference demos, and company announcements. These sources include an article comparing Google's Tensor Processing Unit with NVIDIA's machine-learning focused processors,[106] NVIDIA's resources for deep learning research,[107] and recent announcements from Amazon[108] and Intel.[109] These articles also included qualitative comparisons between current hardware components and those scheduled to be released in the next year.

Tracking NeurIPS author affiliations

In order to catalog and track the author affiliations of all NeurIPS papers from 1987 to 2018, we wrote a Python program to extract author affiliations and assign them to geographic locations. Searching for affiliations required finding the first lines in a paper that use one word from a set of terms, including "university," "college," or "lab." The program then performs a set of text processing steps to consistently format the author's potential affiliation; this includes transforming

102. "Comparison of AI Frameworks | Skymind" 2019; "8 Best Deep Learning Frameworks for Data Science enthusiasts" 2018.
103. "Comparison of AI Frameworks | Skymind" 2019.
104. Ibid.
105. "Open sourcing Sonnet - a new library for constructing neural networks | DeepMind" 2019; "DeepMind hopes its TensorFlow lib Sonnet is music to ears of AI devs - The Register" 2019.
106. "Cost comparison of deep learning hardware: Google TPUv2 vs Nvidia Tesla V100" 2019.
107. "AI Research & Development | NVIDIA DGX Systems" 2019.
108. "Announcing AWS Inferentia: Machine Learning Inference Chip" 2019.
109. "Cheaper AI for everyone is the promise with Intel and Facebook's new chip - MIT Technology Review" 2019; "Intel details Nervana, a neural network chip for inference-based workloads (Updated) | VentureBeat" 2019.

the text into lower case, removing non alpha-numeric characters, and performing limited changes to language (for example converting "lab" to "laboratory"). We then successively ran the program while manually adding terms and text processing steps to identify as many affiliations as possible. Once the author affiliation has been extracted and formatted, we perform a Google Places query on the affiliation text in order to assign it to a geographic location; this enables us to extract the author's home country. In cases where the query failed to automatically assign a geographic location, we manually updated the author's country.

This program was able to identify potential affiliations for 7,099 papers in the original dataset of 8,250 NeurIPS papers. A total of 1,279 distinct author affiliations were found. However, due to irregular formatting or errors in converting the PDF papers to text, some of the affiliations of the 7,099 papers may have been misidentified. Of the 1,279 affiliations, 393 were found to publish two or more papers.

Chapter 3

Past methods for limiting dual-use technology proliferation

Chapter 2 established how the components of deep learning systems have become more accessible and geographically decentralized. It also examined how anonymity and open-sourcing characterize deep learning proliferation. This chapter studies how policymakers have mitigated risks from other dual-use technologies, including nuclear weapons, rocket engines, and biological agents. In particular, we discuss how norms, limiting supply, and controlling exports have been used to reduce dual-use risks. This enables us to develop comparisons between deep learning and other technologies and to analyze how existing models of limiting dual-use technology proliferation could yield relevant lessons.

This chapter begins by asking why governments choose to regulate technology. In particular, we examine how regulators rely on the process of technology assessment to evaluate environmental and social consequences. Although technology assessment can yield important insights about dual-use risks and environmental hazards, past examples indicate that governments struggle to accurately predict diffusion or social effects. We subsequently examine three existing methods for limiting dual-use technology proliferation: Norms and taboos, limiting supply, and limiting exports. We begin by analyzing how norms surrounding chemical and nuclear weapons developed and were codified in international law; we then study the US Nuclear Regulatory Commission and Federal Select Agent Program as examples of the US government limiting the supply of dual-use technology. Finally, in

considering the US International Trade in Arms Regulations (ITAR) and Export Administration Regulations (EAR), we examine how the US government controls dual-use military and commercial technology.

This chapter concludes with our argument about deep learning proliferation. We present three hypotheses: Deep learning's unique technological components yield a distinct proliferation process; most existing regulatory regimes are inapplicable to controlling deep learning; and technical countermeasures may be effective in delaying or disrupting malicious deep learning systems.

3.1 Forecasting social, economic, and security risks

3.1.1 The problem facing US policymakers

Regulating deep learning constitutes one application of a fundamental problem faced by policymakers and commercial actors:

How should highly impactful cutting-edge technology be regulated, and what mechanisms are available for limiting nefarious use?

In this section, we provide an overview of technology assessment, or the process by which public and private actors try to anticipate the effects of technology on society. We then discuss how these assessments drive governments to limit the proliferation or use of a particular technology. After dividing these mechanisms for limiting proliferation into three categories, we discuss why similar policies would not comprehensively limit deep learning applications.

3.1.2 Closing the gap between pure science and application

Technology assessment requires policymakers to analyze how a technology will impact society. As demonstrated by countless recent scientific breakthroughs from electricity and combustion engines to nuclear weapons and transistors, "pure science," which is conducted in laboratories and universities, may yield technology with immense and unforeseeable societal and economic effects. In his 1961 article *On Some Social Consequences of Scientific and Technological Change*, Walter Rosenblith argues that basic scientific research has become increasingly directed towards generating societal impact. This shift in the direction of research from theoretical to practical has increasingly

required innovators and engineers to understand and propose regulations for new scientific developments.[1] Rosenblith studies this hypothesis in the microcosm of MIT course descriptions, noting that course summaries increasingly emphasized opportunities for students to deploy technology "on a large-scale basis for the direct benefit of people."[2] In compelling students to focus on research with an immediate impact, Rosenblith contends that pure science has become directed at the "end product" of solving environmental problems or advancing economic goals.[3] The consequence of this shift: Engineers and scientists now "predict and control" the byproducts of innovation.[4]

Rosenblith's thesis encapsulates the underlying and increasing difficulty of technology assessment as well as the motivation for government action. By the mid twentieth century, technology had yielded new devastating weapons while both creating and alleviating environmental problems, such as air and water pollution.[5] Thus, given this ability to yield both positive and negative consequences, assessing how new technology will impact society has become an important responsibility for both policymakers and engineers. Rosenblith's article holds particular relevance to contemporary deep learning research. Unlike some areas of scientific research, the latest deep learning papers are routinely "exploited for social purposes," from supporting medical devices to improving self-driving cars.[6] As a result, new innovations in deep learning yield "large-scale" societal impact that warrants thorough technology assessment. In the following section, we provide a basic overview of technology assessment and its role in motivating regulation.

3.1.3 Technology Assessment

Why does the government regulate technology? Over the last century, legal textbooks and journal articles have used the context of world-changing scientific developments - such as the creation of the automobile, the development of nuclear fission, and the miniaturization of the transistor - to answer this question. In increasing productivity and access to information, technology is widely perceived as having universally increased human standards of living.[7] Yet, technology also has created new societal problems on a massive scale, including job loss and environmental pollution. When new

1. Rosenblith 1961, 513.
2. Ibid., 502.
3. Ibid., 513.
4. Ibid., 513.
5. Ibid., 502.
6. Ibid., 499.
7. Ibid., 498.

scientific breakthroughs are discovered, how can governments anticipate their impact?

As new technology may undermine common welfare and national security, regulation can be an effective tool for limiting economic or security risks. In *Channeling Technology Through Law* (1973), Laurence H. Tribe presents two fundamental goals for government-led technology assessment: Measuring environmental impact, including the consumption of resources and adverse effects on the environment; and measuring social impact, such as job loss. Forecasting, or predicting the diffusion of technology in society, requires governments to measure social or environmental change. However, in citing the unprecedented proliferation of television receivers throughout the US in the late 1940s and early 1950s, Tribe emphasizes that forecasting may be impossible, or rooted in faulty assumptions about adoption or innovation.[8] As technology may change significantly during the diffusion process, initial assessments or regulation can limit politicians' ability to address "problems that are essentially speculative and remote in time."[9] Ill-conceived regulation may also stifle innovation or diffusion. Tribe's evaluation suggests that understanding a technology's propagation in society is a key prerequisite for anticipating social or environmental effects.

Analyzing the diffusion of innovation

In *Diffusion of Innovations* (1962), Everett Rogers provides a series of concrete questions for policymakers to evaluate a technology's diffusion process. Rogers proposes five characteristics of an innovation that affect its spread through society: Its relative advantage (how much better is it), compatibility (does it fit adopters' needs), complexity (how difficult is it to use), trialability (can it be experimented with), and observability (is this innovation visible to others). Rogers continues to argue that the key enablers and predictors of diffusion include the innovation itself, the communication channels used to share it, the time it takes for diffusion to occur, and the social system surrounding the innovation.[10] Thus, in order to control or regulate technology, the government may regulate each element of diffusion. For example, regulations may restrict communication channels (using import or export controls). Chapter 2 provides data for evaluating Rogers' five characteristics in the context of deep learning. In particular, deep learning's rising trialability, decreasing complexity, and increasing relative advantage may explain its rapid diffusion.

8. Tribe 1973, 29-30.
9. Ibid., 31.
10. Rogers 1962, 35.

Mechanisms for government control

In *Channeling Technology Through Law*, Tribe presents three mechanisms for governments to influence diffusion and technological development: Issuing directives, modifying market incentives, and changing decision-making structures.[11] While directives may specify intellectual property ownership or regulatory limitations, altering market incentives includes imposing taxes or public subsidies. Tribe discusses how creating public agencies can be used to change decision-making structures influencing the diffusion of technology; for example, establishing an agency responsible for licensing uranium mining will inevitably alter how nuclear energy companies approach new projects. The specific case of uranium licensing is discussed further in this chapter.

3.1.4 Methods for mitigating dual-use risks

Rogers, Tribe, and Rosenblith provide a clear structure for understanding how and why regulations evolve for limiting the uses of certain technology. Literature on regulation indicates that there are four ways governments have attempted to control, limit, or counteract emerging technology:

1. **Establishing norms and taboos**, such as those surrounding nuclear weapons, chemical attacks, and landmines;

2. **Limiting supply**, such as the Nuclear Regulatory Commission's licensing requirements for radioactive material;

3. **Regulating exports**, such as the International Trade in Arms Regulations (ITAR) and Export Administration Regulations (EAR); and

4. **Developing technological countermeasures**, such as the Strategic Defense Initiative or gas masks.

This chapter examines the first three of these methods, or how governments have applied norms and taboos, limited supply, and used export controls to restrict technological development and diffusion. Chapter 5 studies how technological countermeasures could be applied to control deep learning.

11. Tribe 1973, 52.

In the remainder of this chapter, we rely on Tribe's framework for technology assessment and Rogers' study of the diffusion of technology to analyze these three types of restrictions. In all three cases, we discuss how existing attempts at technology regulation could be applied to regulate deep learning. However, we ultimately conclude that deep learning's unique proliferation process among millions of anonymous actors indicates that norms, supply controls, and export controls are insufficient to effectively impede malicious uses or limit proliferation.

3.2 Norms and taboos

Scott Sagan defines norms as "shared beliefs about what actions are legitimate and appropriate in international relations."[12] Norms encourage states or other actors to abide by certain standards by enforcing reputational, economic, or other costs on deviation.[13] We begin by analyzing how the stigmatization of chemical weapons as "morally illegitimate" by politicians and individuals led to their exclusion from conventional warfare in World War II and other conflicts.[14] Then, we discuss how other mechanisms were used to develop robust international norms for nuclear weapons in the twentieth century. We question how these technologies - chemical weapons, nuclear weapons, and biological agents - could yield insights for limiting deep learning proliferation and conclude that deep learning's anonymity and relatively unmitigated proliferation on the internet indicates that enforcing norms for individuals would be untenable. However, norms may be effective in altering institutional research practices by companies or universities.

3.2.1 Chemical weapons

Norms proved particularly powerful in changing behavior and motivating international treaties governing chemical weapons. Chemical weapons (CWs), including chlorine, phosgene, and mustard gas, were widely used in World War I.[15] Although CWs killed only 90,000 soldiers during the war, an additional 1.3 million men were blinded, disfigured, or debilitated by CWs.[16] Public outrage prompted politicians and regular citizens to push for their exclusion from conventional warfare.[17]

12. Sagan 1997, 73.
13. Martinsson 2011.
14. Price 1995a, 73.
15. Fitzgerald 2008.
16. "Fact Sheet 1 - History" 2019.
17. Price 1995b, 74.

Today, chemical weapons have been rejected as morally illegitimate and banned by international treaties, including the 1925 Geneva Protocol and the 1993 Chemical Weapons Convention.[18] The 1925 Geneva Protocol banned the first use of biological and gas weapons in warfare, thereby engraving stigmas and taboos associated with CWs into international law. However, countries were not prohibited from developing increasingly potent chemical agents, and many, including the US and European powers, continued to do so during the twentieth century.[19]

Noting that CWs were widely used in major battles in World War I (WWI) but not World War II (WWII) or the Gulf War, historians have traced the origins of CW taboos to the interwar period. Richard Price argues that three factors have generally been used to explain the development of CW norms and taboos since WWI: The possibility of retaliatory action, which could involve the use of CWs against noncombatants; the moral abhorrence of using CWs; and decreasing military preparedness for performing or defending against chemical attacks.[20] Price maintains that the fear of retaliation significantly contributed to nonuse during WWII and the Gulf War.[21] In the 1940s, German and British officers both expressed concern that using CWs on the battlefield could trigger massive chemical attacks against cities or other civilian populations; as a result, the British resisted preparing CWs for doomsday scenarios in the early 1940s and the Germans chose not to use CWs during the Normandy invasion.[22] British leaders also explicitly expressed moral qualms about using CWs against enemy combatants.[23] Similarly, Price suggests that, during the Gulf War, Saddam Hussein hesitated to use CWs because Bush administration officials viewed them as a "red line" that could fundamentally change the conflict and trigger overwhelming retaliation potentially including the use of nuclear weapons.[24] As a result, a combination of morality and fear of retaliation prompted states to pause before utilizing CWs.

Despite the stipulations of the 1925 Geneva Protocol, international norms and taboos have not wholly eradicated CW's use in armed conflict. From 1935 to 1936, Mussolini's armies employed mustard gas against Abyssinian troops in Ethiopia,[25] and, during the Iran-Iraq War, Saddam

18. Beaumont 2018.
19. Prohibition of Chemical Weapons+ 31 70 416 3300 http://www. opcw. org 2015.
20. Price 1995b, 74.
21. Ibid., 77.
22. Ibid., 76.
23. Ibid., 77.
24. Ibid., 77.
25. Grip and Hart 2009.

Hussein used mustard gas and neurotoxins to kill tens of thousands of Iranian troops as well as tens of thousands of Kurds in the Anfal genocide.[26] Though the 1925 Geneva Protocol still allowed states to develop chemical agents, the 1997 Chemical Weapons Convention (CWC) banned the "development, production, stockpiling, transfer and use of chemical weapons."[27] Yet, despite Syria's 2013 accession to the CWC, Syrian President Bashar al-Assad has continued to use CWs against civilians and enemy combatants, further suggesting that norms on conventional warfare cannot completely change behavior, even when codified in international law. Although these transgressions appear to reflect norms' impotence, some - including Price - view these violations as contributing to the "substantial strengthening"[28] of the anti-CW norm by triggering international opposition, such as President Trump's decision to order a cruise missile strike on Syria following a chemical attack.[29]

Could CW norms and treaties be applied to deep learning?

This stigmatization of CWs, which began with the technology's branding as morally abhorrent and culminated in formal bans in international treaties, reveals clear parallels to deep learning. As with CWs, the creation of embarrassing or pornographic deep fakes has been deemed unethical and morally unconscionable by politicians and journalists.[30] This general recognition of the technology's detrimental impact could prompt states to agree on norms for development and use. However, unlike CWs, which require significant scientific research and industrial manufacturing capabilities, deep fakes can be generated by almost any individual with access to the internet. Additionally, while CWs killed or disfigured millions, deep fakes have not yet caused a national security crisis or triggered public outrage. States hesitate to use CWs due to the possibility of retaliation, moral concerns, or military unpreparedness. In contrast, retaliating against or imposing moral norms on anonymous deep fake creators appears impossible. Thus, even though states may agree upon norms for deep learning or deep fakes, changing the the behavior of distributed anonymous actors is untenable. However, norms could change research practices by companies and universities. For

26. "Timeline of Syrian Chemical Weapons Activity, 2012-2019 | Arms Control Association" 2019; "Anfal: Campaign against the Kurds" 2007.
27. "Fact Sheet 1 - History" 2019.
28. Price 1995a, 98.
29. Borger and Beaumont 2018.
30. D. Lee 2018b.

example, these institutions could be encouraged or required to consider the security consequences of open-source research projects or deep learning papers. Norms for open-sourcing are discussed in more detail in Chapter 6.

3.2.2 Nuclear weapons

International norms governing nuclear weapons (NWs) provide a different lens for understanding how taboos and regulations develop. Nina Tannenwald argues that a half-century of domestic and international pressure contributed to the development of a taboo against the first use of nuclear weapons; Tannenwald describes how this norm developed well before the Nonproliferation Treaty (NPT) prohibited non-nuclear weapons states from acquiring the bomb. In "Stigmatizing the Bomb: Origins of the Nuclear Taboo," Tannenwald argues that rationalist political theory about deterrence is insufficient to explain nuclear taboos. Instead, military and civilian leaders were influenced by a combination of public opinion, deterrence, morality, and international politics. These factors ultimately led to NWs' exclusion from conventional warfare.[31] Tannenwald emphasizes that the US and other NW states defied pressure and advocated for their right to first use; however, bottom-up advocacy from social groups, pressure from non-NW states, and ethical principles prompted leaders to submit to international norms. Citing President Truman's enforcement of civilian control over the bomb through the US Department of Energy, Tannenwald continues to argue that norms changed leaders' "identity and interests."[32] Despite later efforts by the military to "conventionalize" tactical NWs, public opinion remained firmly against first use and contributed to the signing of the Partial Nuclear Test Ban Treaty in 1963 - the first time weapons testing had been restricted by international law.[33]

Over the following half-century, public opinion, changing institutional priorities, and leaders' increasing preference for restraint solidified NW norms. As the use of NWs was stigmatized, other institutions, including the NPT and the International Atomic Energy Agency, began to advocate for the peaceful use of nuclear energy. In her book "The Nuclear Taboo: The United States and the Non-Use of Nuclear Weapons," Tannenwald emphasizes that - unlike nuclear weapons - the NPT's

31. Tannenwald 2005, 5.
32. Ibid., 14.
33. Ibid., 27.

treatment of nuclear energy enforces "equality among states."[34] Thus, as the pursuit of nuclear energy was permitted while nuclear weapons were banned, norms and treaties separated peaceful and malicious uses of nuclear technology.

Applying NW norms to deep learning

In contrast to nuclear weapons in the late twentieth century, public opinion on deep learning and deep fakes has not crystallized to the extent necessary to receive significant attention from world leaders. As NWs were used to kill over one hundred thousand individuals in WWII, were tested frequently in the 1950s, and were employed as deterrents throughout the entire Cold War, public pressure on international leaders increased significantly. In contrast, although deep fakes are widely accessible on the internet, they have not caused mass-casualty events or precipitated national security crises.

It is also critical to emphasize how NW norms restricted behavior for a highly complex technology and limited number of actors. When the NPT was signed in 1968, only five states had tested nuclear weapons.[35] Furthermore, when nuclear weapons norms are breached, such as in an illegal test, investigators may use nuclear forensics to identify and shame the responsible actor. In other cases, violations can be discovered from the enormous machinery, including reactors, centrifuges, and biohazard equipment, necessary to construct nuclear weapons. In contrast, any individual with internet access can create or publish deep fakes. As deep learning research and applications are distributed across millions of actors (many of whom are anonymous), creating or enforcing "shared beliefs about what actions are legitimate and appropriate in international relations" may be impossible.[36]

In prohibiting nuclear weapons while encouraging nuclear power, the NPT also yields lessons in selectively developing norms around different uses of similar technology.[37] Deep learning can yield societal benefits in medical technology, transportation, and other industries. Thus, just as the NPT preserved countries' rights to pursue nuclear energy, perhaps norms could develop around applying deep learning for peaceful purposes while preventing its use in surveillance technology or malicious

34. "Treaty on the Non-Proliferation of Nuclear Weapons (NPT) - UNODA" 2019; Tannenwald 2007, 335.
35. "Nuclear weapons timeline | ICAN" 2019.
36. Sagan 1997, 73.
37. "Treaty on the Non-Proliferation of Nuclear Weapons (NPT) - UNODA" 2019.

applications.

3.3 Limiting supply

Limiting supply, or regulating access to the components required to produce hazardous technology, has been widely used by the US government to limit dual-use technology proliferation. In managing risk by controlling production, supply controls have been applied by governments to particularly sensitive technology, such as fissile material and chemical or biological agents. We examine how supply controls evolved to prevent the misuse of nuclear material and biological agents. We then analyze how similar supply controls could be applied to the technological precursors of deep learning: Data, software, and hardware.

3.3.1 Regulating nuclear material

Under Laurence Tribe's delineation of mechanisms for controlling technology, the US Nuclear Regulatory Commission (NRC) enforces specific directives and changes decision making structures. By imposing strict mining and licensing requirements on fissile materials, the NRC regulates nuclear "source material," which can be used to produce a nuclear weapon.[38] This includes all civilian uses of source material, such as uranium mining, nuclear reactor operation, medical and academic uses, and radioactive waste management. Companies seeking to refine uranium ore, to enrich uranium for nuclear fuel, or to process spent fuel must receive a license from the NRC.[39] [40]

Licensing requirements for constructing new nuclear reactors, operating nuclear reactors, and disposing of nuclear waste are specified with excruciating detail, and all proposed reactor facilities are subject to engineering and environmental reviews.[41] Facility applicants must also outline how proposed safety features would prevent specific accident scenarios, such as reactor meltdowns and radiation exposure. Under Tribe's delineation of technological development regulation, this process of collaborative development and licensing can be categorized as a modification of the decision making structure in the nuclear power industry; scientists, engineers, and operators must work

38. "NRC: Source Material" 2019.
39. "NRC: Medical, Industrial, & Academic Uses of Nuclear Materials" 2019.
40. Other domestic bodies and international agencies, such as the International Atomic Energy Agency, regulate nuclear material and ensure compliance internationally; we focus specifically on US regulations for fissile material.
41. "NRC: Backgrounder on Nuclear Power Plant Licensing Process" 2019.

with the NRC to exploit nuclear power.

The NRC's evolution over time

The NRC's structure and regulatory requirements have evolved significantly since the organization's inception in 1975, particularly following the Three Mile Island partial reactor meltdown that occurred in 1979.[42] Although the organization has always tightly controlled access to fissile material, the 1979 incident prompted the NRC to create more resources for crisis response, increase the frequency of plant inspections, and institute new reporting requirements.[43]

Recent NRC performance reports, which track the unintentional release of nuclear material, unhealthy radiation exposure, and other safety metrics, indicate that US nuclear plants have been operating within desired standards.[44] However, despite these statistics and the strict procedural requirements outlined on the NRC's website, government accountability reports have suggested that many nuclear facilities operate outside of licensing limitations, and inspectors have given facilities flexible timelines for resolving safety concerns.[45] One post-9/11 government accountability report also maintained that nuclear power facilities lacked sufficient perimeter security, inspection programs, and crisis response strategies to potential terrorist attacks.[46]

The NRC and cutting-edge technology

The NRC's efforts to keep up with innovation in the nuclear industry offer additional insights into the challenges associated with tightly controlling supply. Although the NRC has published plans to update guidelines to accommodate more advanced reactors, commercial developers have expressed skepticism that the commission will be adequately prepared to license novel nuclear energy technology.[47] As a result, some reports have suggested that startups developing nuclear power technology may pursue construction or licensing opportunities outside the US.[48] These shortcomings illustrate potential pitfalls of government modifications to industry decision-making structures: When introducing bureaucracy and requiring government input, commercial projects may face obstructions.

42. Temples 1982.
43. "NRC: Backgrounder on the Three Mile Island Accident" 2019.
44. "NUREG-1542, Vol.21, Suppl 1, "Fiscal Year 2015, Summary of Performance and Financial Information."" 2019.
45. Jones 1998.
46. *Oversight of Security at Commercial Nuclear Power Plants Needs to Be Strengthened* 1998.
47. "NRC Vision And Strategy For Licensing Advanced Reactors Needs Improvement" 2019.
48. Commission et al. 2016.

In the following section, regulation's deleterious effects on innovation are further explored in the context of biological agents.

3.3.2 Regulating the machinery and supplies for biological agents

US government regulations for hazardous toxins and biological agents present another example of how specific directives and decision making structures evolved to control supply and mitigate technological risk. The Federal Select Agent Program (SAP), which is jointly administered by the Centers for Disease Control (CDC)[49] and the Department of Agriculture, is tasked with overseeing and controlling the possession and transfer of toxic chemicals and select agents.[50] SAP currently regulates 67 select agents and toxins, which include the variola major virus (smallpox), the yersinia pestis virus (plague), and bacillus anthracis (anthrax). Although initially created in response to an illegal and unregulated dispersal of plague virus in 1995, SAP regulations were vastly expanded by the 2001 Patriot Act and 2002 Public Health Security and Bioterrorism Preparedness and Response Act following 9/11 and the 2001 anthrax attacks. This legislation expanded the number of regulated agents and increased the federal government's tracking and oversight capabilities.[51] Additional requirements are imposed on facilities that work with particularly sensitive agents, such as smallpox; these select agents, which are perceived as having the "greatest risk of deliberate misuse" or "significant potential for mass casualties or devastating effect to the economy, critical infrastructure, or public confidence," are designated as "tier one" select agents and require additional safeguards for processing or use.[52]

SAP regulations have appreciably increased federal oversight and monitoring of particularly dangerous agents. However, several widely publicized incidents where tier one agents (including anthrax) were improperly handled have prompted concern over the quality of SAP inspections and security protocols.[53] Internal reviews, government accountability reports, and industry researchers have noted other shortcomings. One glaring issue: Many select agents can be synthesized with relative ease or smuggled into the US in small quantities, thereby undermining SAP protections.[54]

49. The CDC is part of the US Department of Health and Human Services.
50. "Federal Select Agent Program - About Us" 2019.
51. "Public Health Security and Bioterrorism Preparedness and Response Act of 2002 | Department of Energy" 2019.
52. "Select agents & toxins FAQ" 2019.
53. "Report of the Federal Experts Security Advisory Panel - December 2014" 2019.
54. House 2013.

Some studies have suggested that extremely harmful pathogens, such as ebola, could be imported and completely circumvent SAP protections. Furthermore, due to increasingly strict security requirements associated with tier one select agents, fewer scientists now conduct research on these particularly sensitive toxins, and collaborating with international partners or hiring researchers has become complicated for US laboratories. Thus, beyond using licensing requirements to modify the decision-making processes used by laboratories, SAP also has altered market incentives by significantly increasing the costs associated with research. These requirements have prompted some researchers to relocate abroad or refocus their efforts.[55] Other scientists have criticized the SAP's accounting system as unsuited for storing biological agents. As government administrators may lag behind in rectifying these regulatory limitations or failures, SAP's drawbacks suggest that strict licensing, inspection, and safety requirements could refocus or impede scientific research.

3.3.3 Applying supply controls to deep learning

Could limiting supply reduce deep learning risks in a manner similar to radioactive material or select agents? As outlined in Chapter 2, training deep learning models requires datasets, software libraries, and computing hardware.

Regulating data: Deep learning's uranium?

Restricting access to the data required for deep learning poses significant complications and drawbacks. Limiting access to data could be implemented through regulation that specifies the types of information that can be shared or downloaded online. Alternatively, following the model of the NRC's oversight of all nuclear material operations, regulators could require websites such as Kaggle, which shares datasets, to receive licenses to distribute certain types of data, such as bulk images of faces. Within the last year, researchers at Microsoft have called for the government to regulate facial recognition, including the collection of training images.[56] This proposal could be implemented in a manner similar to the SAP for biological agents or NRC for nuclear material. However, unlike nuclear source material, which is tangible, measurable, and detectable, datasets can be shared anonymously on the internet. Once individuals have been given unfettered access

55. "Fast Track Action Committee Report : Recommendations on the Select Agent Regulations Based on Broad Stakeholder Engagment - October 2015" 2019.

56. Smith 2018.

to data, they can effortlessly and anonymously share it. Thus, the accessibility and diffusion of datasets poses challenges that literature on technology regulation fails to address.

Limiting access to software

Limiting access to open-source software libraries would also prove impractical due to their diffusion through the internet. In some cases, governments and regulatory agencies have struggled to track the materials required to build nuclear or biological weapons. For example, after Pakistani scientist A.Q. Khan stole plans for Urenco's L-1 and L-2 gas centrifuge, an estimated thirty companies in twelve countries gained access to sensitive information on developing nuclear weapons.[57] However, unlike gas centrifuges and uranium ore, which are already difficult to track, the open-source software used as building blocks for deep learning are distributed among *orders of magnitude more* actors. While building enrichment facilities requires significant financial and logistical resources, training deep learning models can be done at little expense with access to the internet.

Thus, even if access to official code repositories were restricted from now onward, individuals could easily locate other sources for downloading existing models. However, it may be possible to introduce mechanisms that track deep learning models' dissemination, such as licensing or signup requirements on websites. Individuals may be able to fraudulently circumvent such tracking programs, and limiting scientists' ability to access or use certain models could significantly hamper deep learning research in the US, just as the security protocols imposed by SAP have impaired researchers' ability to experiment on certain biological agents. We discuss these proposals in greater detail in Chapter 6

The NRC's struggle to regulate novel reactor technology suggests other issues associated with limiting access to software. While policymakers could design regulations for existing technology, unconventional systems designed in future years could pose additional difficulties or introduce new risks. As a result, although the literature suggests that limiting supply can effectively reduce risks for some dual-use technology, these mechanisms present clear limitations for controlling access to open-source software.

57. "Companies Reported to Have Sold or Attempted to Sell Libya Gas Centrifuge Components | NTI" 2005.

Limiting access to hardware

Similar to the NRC's restrictions on uranium refinement and nuclear power plants, deep learning regulations could restrict individuals' ability to purchase or rent training hardware, such as graphics processing units (GPUs) or tensor processing units (TPUs) capable of quickly performing matrix calculations. However, regulating deep learning hardware also appears impractical. Multiple GPU manufacturers, such as Asus, are based outside of the US, and thousands of varieties of GPUs are available for purchase and shipment across myriad internet vendors.[58] Furthermore, although advanced GPUs and TPUs may increase training speed by multiple orders of magnitude, it is still possible to train deep learning models on older, less sophisticated computing hardware, such as the CPU in a personal laptop.[59][60] As a result, controlling access to computing hardware would likely be futile in comprehensively preventing malicious deep learning applications. In some research applications, training models may still require weeks of computing time on highly efficient systems. Thus, should certain types of training hardware eclipse current TPU or GPU training speeds by additional orders of magnitude, government controls restricting the sale or use of these highly efficient devices could limit individuals' ability to design and test deep learning models. Additionally, researchers have begun to explore options for training particularly large deep learning models on thousands of parallel machines. Because access to hardware is particularly important for training these complex models, restricting or monitoring access to hardware could yield valuable information for policymakers.[61] This case further indicates how deep learning necessitates a different approach to counter-proliferation that focuses on delaying or disrupting malicious systems; although technical countermeasures may play a key role in prevention, other mechanisms, such as tracking hardware usage, may also yield valuable insights.

58. "About ASUS - Facilities & Branches" 2019.
59. Differences in training on CPU and GPU hardware are explored further in Chapter 2.
60. "TensorFlow performance test: CPU VS GPU - Andriy Lazorenko - Medium" n.d.
61. Goyal et al. 2017.

3.4 Export controls

Motivated by the dangers posed by sensitive technology spreading to rival states or rogue actors, US policymakers have evolved sophisticated regimes for controlling exports of dual-use technology. The International Trade in Arms Regulation (ITAR) and Export Administration Regulations (EAR) are extensive legal regimes created by the US government to prevent military equipment and commercial dual-use technology from spreading to foreign nationals. In this section, we begin by discussing how ITAR proactively limits foreign persons' access to military technical data while EAR prevents exports of dual-use technology; under Tribe's framework in *Channeling Technology Through Law*, both regimes represent specific directives used by the government to manage the diffusion of innovation and associated security risks.

3.4.1 The International Trade in Arms Regulation (ITAR)

ITAR prohibits individuals and companies from exporting sensitive defense articles outside of the US. In particular, ITAR limits the transfer of blueprints, software, and other "technical data" associated with rockets, nuclear weapons, aircraft, tanks, military training devices, and numerous other categories of defense equipment.[62] The term "technical data" is applied liberally: Any physical or digital files related to controlled technology may be protected under ITAR. Thus, to ensure compliance, companies may rely on encryption and data classification schemes (where any sensitive documents are digitally encrypted or physically marked with "export controlled" designations).[63] Some researchers and engineers have argued that these restrictive regulations and bureaucratic processes have reduced the competitiveness of US companies abroad as they may be unable to collaborate with some foreign entities.[64]

3.4.2 ITAR and deep learning

ITAR's reliance on bureaucratic licensing and security protocols to keep sensitive information inside the US does not provide a particularly useful model for deep learning. While it may be logical to restrict exports of rocket engine and missile blueprints, the core theoretical and technical building

62. "22 CFR Subchapter M - INTERNATIONAL TRAFFIC IN ARMS REGULATIONS | CFR | US Law | LII / Legal Information Institute" 2019.
63. Hatmaker 2017.
64. Tushe 2011.

blocks of deep learning are already open-source and distributed internationally. Thousands of academic papers, millions of lines of code, and numerous open-source libraries have been developed and used in companies and research labs around the world. ITAR's impact on the competitiveness of US companies abroad reveals additional considerations for deep learning. While ITAR primarily protects devices with strictly military applications (such as cruise missiles), deep learning has already led to promising breakthroughs in healthcare and autonomous vehicles. Furthermore, although US companies and universities may currently lead in deep learning research, domestic restrictions could prompt a "brain drain" to other countries.

Recent events suggest that ITAR does yield some lessons for limiting deep learning models' diffusion. In February 2019, OpenAI released findings from its GPT-2 language model, which is able to generate coherent paragraphs of text on particular topics.[65] Noting potential "concerns about malicious applications of the technology," OpenAI chose to only release a subset of the entire GPT-2 model; released versions of the model have 117 and 345 million parameters while the full, unreleased model has 1.5 billion parameters. More recently, OpenAI announced they would share the model with other AI and security researchers.[66] Government agencies could conceivably supersede this commercial decision-making authority and impose regulations on the types of models that can be open-sourced. In this case, OpenAI recognized how "malicious actors—some of which are political in nature—have already begun to target the shared online commons;"[67] other companies or individuals may not consider these consequences. Export controls could limit companies from releasing certain types or sizes of deep learning models; we study this proposal and other policy implications in Chapter 6.

3.4.3 Export Administration Regulations: Limiting dual-use risks

While ITAR focuses on keeping defense-related technical data inside the US, EAR are primarily concerned with exports of commercial dual-use technology. EAR divides regulated technology into ten categories, which include "Nuclear," "Computers," "Information Security," and "Materials, Chemicals, Microorganisms and Toxins."[68] Each category references a lengthy document that

65. "Better Language Models and Their Implications" 2019.
66. "openai/gpt-2: Code for the paper "Language Models are Unsupervised Multitask Learners"" 2019; "Better Language Models and Their Implications" 2019.
67. "Better Language Models and Their Implications" 2019.
68. "Commerce Control List (CCL)" 2019.

specifies subcategories of controlled technology. Export controls on software products are particularly sensitive to whether encryption algorithms are used.[69] The "Computers" category specifies additional export controls for software and hardware. For example, some software that performs "real-time processing" may be export-controlled.[70] While both ITAR and EAR enumerate specific regulations for different categories of sensitive technology, EAR provides more detailed regulations on software products.

Applying EAR to deep learning

Could the government rely on similarly specific controls for deep learning? Given EAR's licensing requirements for software based on level of encryption, neural networks, which are used to perform deep learning, could be regulated with similar specificity. For example, EAR could require mandatory export controls based on the number of parameters in a neural network. Just as OpenAI released a smaller version of their 1.5 billion parameter network, perhaps regulations could prohibit entities from releasing types of models based on the number of parameters. However, as with nascent regulations for nuclear power and biological agents, these guidelines could prove inflexible and impede innovation. They may also become obsolete as new models are proposed. We explore these proposals further in Chapter 6.

3.5 Technical countermeasures: An effective alternative?

While it may be possible for corporations and governments to develop norms of behavior, anonymous actors pose challenges for enforcement. Supply and export controls are also present clear disadvantages for preventing individuals or rogue agents from developing malicious deep learning applications; these mechanisms for limiting proliferation could also stifle innovation in deep learning research. Technical countermeasures, or technology designed to limit proliferation or undermine deep learning systems, may yield effective alternatives. DARPA has recently awarded contracts to private companies to develop tools that detect deep fakes; theoretically, these systems could be provided to companies or government agencies to enable greater information security.[71] Computer

69. "Export Administration Regulations (EAR)" 2019.
70. Ibid.
71. "DARPA is funding new tech that can identify manipulated videos and 'deepfakes' | TechCrunch" 2019.

science research groups around the world have also begun to publish papers outlining algorithms capable of identifying fake audio and video. However, as these efforts address only one malicious use case of deep learning, a more comprehensive approach is necessary.

3.6 Argument

The background information in Chapter 2 and three mechanisms for controlling proliferation examined in this chapter present clear challenges for policymakers. As deep learning technology becomes easier to access, the first step in mitigating harmful applications is understanding its proliferation process. This thesis provides evidence to support three hypotheses:

1. Deep learning is unique in its reliance on open-source and accessible technology; one implication of this uniqueness is that deep learning proliferates differently from other dual-use technologies.

2. Most existing regulatory mechanisms for controlling dual-use technology are inapplicable to deep learning.

3. Technical countermeasures, including disruption or delay mechanisms, can mitigate deep learning risks.

The background information in Chapter 2 and examples of dual use technologies in this chapter provide evidence to support the first hypothesis. The three regulation mechanisms presented in this chapter - norms and taboos, limiting supply, and controlling exports - suggest how existing methods are largely inapplicable to deep learning, thus supporting our second hypothesis. In exploring deep fakes' proliferation from academia to individuals, Chapter 4 provides additional evidence for this hypothesis. By demonstrating how technical countermeasures could disrupt facial recognition systems and prevent the creation of deep fakes, Chapter 5 supports our third hypothesis about how technical mechanisms could limit proliferation or counter deep learning applications. In sum, our hypotheses demonstrate how policymakers should adjust counter-proliferation strategy to employ an amalgamation of technical and policy mechanisms for limiting malicious use.

Chapter 4

Case study: Tracing deep fake proliferation

4.1 Deep fakes: National security risks

4.1.1 Blurring fact and fiction

> "By blurring the line between fact and fiction, deep fake technology could undermine public trust in recorded images and videos as objective depictions of reality."[1]

In September 2018, US Representatives Adam Schiff (D-CA), Stephanie Murphy (D-FL), and Carlos Curbelo (D-FL) wrote a letter to Dan Coates, the Director of National Intelligence, warning of deep fakes' dire national security consequences. Deep fakes are realistic falsified images, audio, or video created using deep learning models.[2] Deep fakes could showcase fake speeches given by high-level political officials (such as a world leader making a statement), be used in sophisticated fake news, or cause widespread panic (for example, deep fake technology could generate deceptive announcements on missile strikes, diseases, or mass shootings). Beyond consequences for national politics, deep fakes present equally concerning implications for individuals. Criminals could create fake videos to extort money from victims or release deep fakes that cause public embarrassment. Even if these videos are proven to be counterfeit, reputation damage could be permanent. As the world

1. "2018-09 ODNI Deep Fakes letter.pdf" 2019.
2. Vincent 2019c.

increasingly relies on the internet to deliver news and information, deep fakes fundamentally changes expectations of truth and reality: Seeing is no longer believing.[3] Over the last two years, deep fake technology has transitioned from academic papers to open-source technology and downloadable applications.

How did deep fake technology spread from academic literature to open-source models and downloadable software in only two years?

Chapters 2 and 3 established how deep learning proliferates differently from other dual-use technologies and how each technological element of deep learning - data, software, and hardware - spreads. This chapter revisits previous findings in the specific case of deep fakes. As no past literature has examined this technology transfer process, this section constitutes one of the first analyses of deep learning proliferation. We begin by discussing media and political responses to deep fakes. Then, by analyzing a series of published papers used to create contemporary deep fake technology, we establish critical points in the proliferation process that could be targeted by policymakers. This case reveals three key stages in deep fake proliferation:

1. **Academic papers**: The technological precursors to deep fakes were first published in scholarly literature and academic papers.

2. **Open-source software**: Soon after academic papers were published and open-sourced, their models were re-released as open-source software specifically designed to generate deep fakes.

3. **Downloadable applications**: As deep fakes became increasingly popular and publicized, programmers repackaged open-source software as downloadable applications that did not require any programming experience to use. This constitutes the last step in making deep fake technology accessible to inexperienced individuals.

4.1.2 Existing literature on deep fakes

Technical academic papers, think tank reports, and news articles constitute most of the existing literature on deep fakes and their security implications. As described in Representative Schiff's letter, some politicians have recognized that the government could play a proactive role in developing

3. Chesney and Citron 2018a.

technology capable of identifying fake audio and video; other reports and articles, including a study by the Council of Foreign Relations, suggest that Congress could pass legislation regulating deep fakes.[4] Noting that this would require Congress to develop and enforce strict and highly specific requirements on how certain video creation technology could be used, this action would be "unlike anything seen previously" and could impede innovation in the field.[5] Beyond potential Congressional action, DARPA's media forensics division has fostered collaboration among academic and commercial institutions to identify deep fakes.[6] DARPA awarded contracts to private companies for the purpose of developing tools to detect deep fakes; computer science research groups around the world have also begun to publish algorithms for identifying fake audio and video.[7] However, deep fake technology continues to evolve.

4.2 Proliferation stage 1: Deep fakes' roots in academia

Although researchers have studied neural networks since the 1950s,[8] the models required to create sophisticated deep fakes were developed and released far more recently. In 2016, computer science researchers at Stanford released a paper describing one of the key technological grandfathers of deep fakes. The paper - called Face2Face - altered an individual's facial expressions based on a recording of another individual's face.[9] By solving a complex optimization problem that models three dimensional facial positions, this system converted facial expressions in one video to align with the facial expressions and lip movements of a recorded speech or interview. For example, given a recording of a politician, such as President Bush,[10] their algorithm could change Bush's facial expressions to match those of another individual. Although this development marks a key milestone in the development of deep fakes, the Face2Face paper did not rely on cutting-edge deep learning techniques and instead described its algorithms as innovations on existing AI techniques that could support "high-end movie production."

One year later, a team of researchers at the University of Washington applied innovations from

4. "Disinformation on Steroids: The Threat of Deep Fakes" 2019.
5. Ibid.
6. Hatmaker 2018.
7. Li et al. 2018.
8. This history is explored in greater detail in Chapter 2.
9. Thies et al. 2016.
10. The original paper used examples of President Bush, Vladimir Putin, Arnold Schwarzenegger, and Daniel Craig.

Face2Face to develop an algorithm that could synthesize video from audio recordings, thereby marking another critical milestone towards the creation of video deep fakes. After training a deep learning model on hours of video, their algorithm - nicknamed "AudioToObama" - could generate a completely fake video of someone speaking from a pure audio recording.[11] For example, given roughly fourteen hours of video of President Obama giving speeches in the Oval Office, their neural-network based machine learning algorithm could convert a pure audio recording of an Obama speech into a video of Obama delivering the speech "on camera." The AudioToObama paper and demonstration videos prompted widespread media coverage as it showcased how video of important politicians could be synthesized from nothing.[12]

Beyond university research groups, companies also began to harness new machine learning algorithms to synthesize fake video and audio. In late 2016, Adobe unveiled a product called "VoCo." Nicknamed the "Photoshop-for-voice," VoCo allowed users to change or edit phrases in audio recordings.[13] Adobe's algorithm required only twenty minutes of an individual speaking in order to convincingly synthesize new phrases.[14] Within months, Alphabet's subsidiary DeepMind, which has pioneered deep learning technology, demoed a similar product called "WaveNet," which used neural networks to "generate speech which mimics any human voice."[15] WaveNet and Adobe VoCo prompted significant media scrutiny and criticism from cybersecurity experts, who stated that voice-mimicking technology could render voice biometric security used by banks and other companies useless.[16] Furthermore, as controversial recordings of President Trump had recently prompted public backlash prior to the 2016 election, other experts wondered whether deep fake technology could also be used to obscure embarrassing content.[17] Given this political context, WaveNet and VoCo represent some of the earliest connections between deep learning and national security. Both of these products, pioneered by tech giants Adobe and Google, also demonstrate how machine learning innovation stems from collaboration between academia and industry. This presents another unique challenges for restricting deep learning proliferation.

At this point, deep fake technology remained squarely in academia and commercial research labs.

11. Suwajanakorn, Seitz, and Kemelmacher-Shlizerman 2017.
12. Vincent 2019b.
13. "#VoCo. Adobe MAX 2016 (Sneak Peeks) | Adobe Creative Cloud - YouTube" 2019.
14. "Adobe Voco 'Photoshop-for-voice' causes concern" 2016.
15. Oord et al. 2016.
16. Armerding 2017.
17. Mak 2018.

Although it was first demoed in 2016, Adobe VoCo has not been commercially released. Google WaveNet was used to support better text-to-speech synthesis but not released as standalone tool for imitating any human voice.[18] However, over the following two years, other researchers and engineers applied innovations from Adobe VoCo, WaveNet, and AudioToObama to create easy-to-use deep fake tools that could be used by the general public to generate counterfeit video or audio.

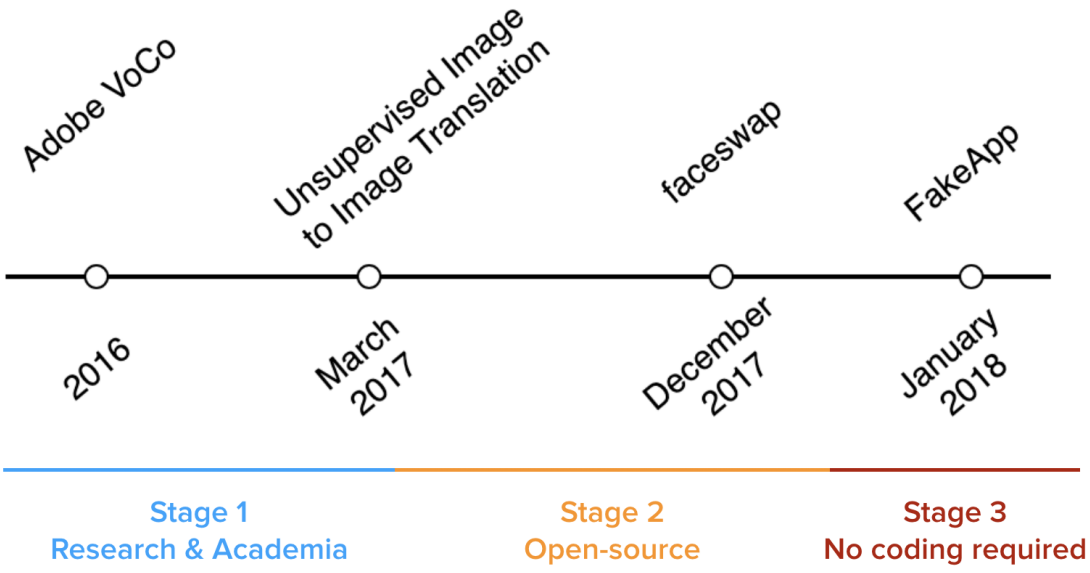

Figure 4.1: Deep fake proliferation stages. Some papers and commercial demos are omitted in this figure but discussed in this chapter.

4.3 Proliferation stage 2: From NVIDIA to individuals: Moving to open-source

Adobe VoCo, WaveNet, AudioToObama, and Face2Face are early examples of how deep learning and artificial neural networks can alter facial expressions, edit phrases in recordings, and generate video from audio. In 2017 and 2018, innovations in deep learning provided the remaining technical tools required to synthesize convincing fake videos of individuals in any context.

In March 2017, researchers at NVIDIA released a paper entitled "Unsupervised Image-to-Image

18. "Cloud Text-to-Speech - Speech Synthesis | Cloud Text-to-Speech API | Google Cloud" 2019.

Translation Networks."[19] Using two recently developed neural network models (variational autoencoders and generative adversarial networks[20]), the authors presented a framework for mapping an image in one setting to an image in another setting. The original paper provided numerous examples of how this technology could be used for sophisticated image-to-image transfer: Animals could be swapped in photos (such as changing a horse to a zebra in a picture of a farm), seasons or weather could be altered (a green field could be transformed into a snowy landscape), or facial attributes could be changed (such as adding facial hair and eyeglasses or completely altering an individual's identity). The models discussed in the NVIDIA image-to-image translation paper are considered to be the theoretical foundation of the latest deep fake technology and were first demoed to the public at NeurIPS (Neural Information Processing Systems) 2017; NeurIPS frequently showcases the latest technology in deep learning and neural networks. After this presentation, NVIDIA released the paper, video examples, and code used to train and test the machine learning models to the general public on the website GitHub, which provides online collaborative tools for sharing and editing code.[21] Any individual with an internet connection could subsequently download and experiment with these models. Although users are required to have a GitHub account to download software, one could easily create an anonymous online persona to gain access.

Because NVIDIA open-sourced the models from "Unsupervised image-to-image translation" after NeurIPS 2017, their paper became one of the key enablers of deep fake proliferation. However, even after the NVIDIA models were released publicly on GitHub in early 2017, creating deep fakes of politicians, celebrities, or other individuals still posed an enormous challenge for programmers without significant experience in machine learning. Working with complex neural networks is computationally expensive and time consuming; furthermore, although the NVIDIA models could perform image-to-image translation in a variety of contexts (for example on images of faces, animals, or landscapes), each specific application requires an extensive training dataset and tuning parameters on the model's learning behavior. Additional open-source software released in the following months alleviated many of these challenges and represent the next stage in deep learning proliferation. This raises one key question: Why did the authors of the NVIDIA paper decide to open-source their code and models? Did the authors consider the privacy and security

19. Liu, Breuel, and Kautz 2017.
20. Generative adversarial networks are described in greater detail in Section 2.2.2.
21. "mingyuliutw/UNIT: Unsupervised Image-to-Image Translation" 2019.

consequences of releasing deep fake technology to the general public? The NVIDIA paper was likely open-sourced to increase the reproducibility of their research and encourage others to replicate or extend their models in future academic research or other projects. Releasing code and datasets significantly reduces the technical barriers and time required to replicate research. Although some applications, such as special effects, pose few security consequences, perhaps the authors did not realize how this technology could contribute to generating sophisticated deep fakes.

"Unsupervised Image-to-Image Translation" solved many of the challenging theoretical problems required to create deep fakes. However, even with access to the models and code from the 2017 paper, synthesizing counterfeit videos remained inaccessible to amateur programmers as it required collecting data and optimizing models. The GitHub library "Faceswap," which was released by an anonymous author on GitHub in December 2017, provided a vastly simplified framework for performing image-to-image translation on videos of people.[22] Unlike the original NVIDIA image-to-image translation library, which requires users to write code in order to train a new model, Faceswap simply accepts two image datasets, where an individual's face from one dataset is translated onto images in the second one. No additional parameters or data is required.

Figure 4.1: One Faceswap user trained a deep learning model to replace other actors' faces with Nicholas Cage's face. Here, Harrison Ford is replaced by Cage in "Raiders of the Lost Ark."[23]

In May 2018, an anonymous contributor named "Torzdf" added a graphical user interface to the Faceswap application. After this update was published, a user simply had to download Faceswap, run a single line of code to launch the graphical application, and specify two datasets with source

22. "deepfakes/faceswap: Non official project based on original /r/Deepfakes thread." 2019.

and target imagery in order to create a deep fake.[24] Over the last two years, fifty-five different accounts (with names from "LordVulkan" to "facepainter") have added additional code to the Faceswap library to improve functionality and ease of use.

Faceswap's attmept to control open-source software

Interestingly, the Faceswap code repository provides a warning to all users:

> "We are very troubled by the fact that faceswap can be used for unethical and disreputable things. However, we support the development of tools and techniques that can be used ethically as well as provide education and experience in AI for anyone who wants to learn it hands-on. We will take a zero tolerance approach to anyone using this software for any unethical purposes and will actively discourage any such uses."

Despite this strongly worded letter and "zero tolerance" policy, Faceswap remains publicly available for download without any form of access control. Within the Faceswap application, there are no built-in technical tools that allow the developers to limit the technology's use. Thus, this toothless warning to users suggests that, despite ostensibly ethical goals, it is impossible to control the applications of open-source deep learning technology. In Chapter 3, we analyze how government-led regulatory regimes, such as ITAR and EAR, were developed to target dual-use technology proliferation. However, this warning clearly demonstrates how the lack of controls for open-source software leaves developers few options for limiting use beyond a strongly worded message. As deep fake technology was again repackaged into a Windows application, these warnings failed to effectively control the underlying technology.

The transition from inaccessible academic models ("Unsupervised Image-to-Image Translation") to easy-to-use downloadable software (Faceswap) provides an illustrative example of the unique challenges posed by deep learning proliferation. After academics developed and released groundbreaking models that applied deep learning to a fundamentally new problem, programmers anonymously repackaged these models into open-source software designed for the specific purpose of creating deep fakes. As Faceswap has likely been used to generate many of the tens of thousands of deep fakes now posted throughout the internet (including the Nicholas Cage video), deep fake proliferation highlights the hazards of open-sourcing technology with dual-use capabilities and national security consequences. After NVIDIA researchers posted their code publicly, it became impossible

24. "deepfakes/faceswap: Non official project based on original /r/Deepfakes thread." 2019.

to control its applications. Additionally, as Faceswap was released and maintained by anonymous contributors, it become impractical for anyone, including NVIDIA, to track current users or developers. In the next brief section, we explore how the Faceswap creators attempted to limit the use of their open-source models.

4.4 Proliferation stage 3: Applications for amateurs

In early 2018, other programmers anonymously repackaged Faceswap's underlying models into a downloadable application called "FakeApp." The application has a simple graphical interface and can be downloaded and run on Windows computers, thus lowering the technical bar for creating deep fakes to the point where absolutely no programming - not even a single line of code - is required to produce deep fake videos.[25]

Numerous websites offer FakeApp downloads and tutorials that provide detailed instructions for creating deep fakes. Tens of YouTube videos, Reddit posts, and blogs narrate the step-by-step process for setting up and using FakeApp.[26] Some videos discuss highly technical optimizations for generating deep fakes, including tuning machine learning parameters. Non-technical journalists have experimented with FakeApp, thereby typifying how deep fake technology has become user friendly; one article published by *The Verge* links to a video where Elon Musk is swapped with Jeff Bezos in a public statement.[27] The stages of increasing decentralization of deep fake technology in its transitions from NVIDIA, to Faceswap, and finally to FakeApp reveal complex challenges for controlling the applications of deep learning: Once this technology left academia, it became impossible to control or track. As government officials and journalists recognized deep fakes' political import, internet users transformed abstruse academic models into downloadable software.

4.5 Revisiting legality and counter-proliferation

Faceswap and FakeApp allow individuals with access to computing power to create deep fakes. Although some politicians and reporters remain concerned with privacy and security implications,

25. "How To Install FakeApp - Alan Zucconi" 2019.
26. "deepfakes guide: Fake App 2 2 Tutorial. installation - YouTube" 2019; "FakeApp download links and How-To Guide : GifFakes" 2019.
27. Robertson 2019.

other government officials and technology experts have emphasized that deep fake technology is not an entirely new national security threat; instead, it represents the democratization of decades-old tools used by professional movie editors. In an anonymous interview, one of the creators of the Faceswap library noted that deceased actor Paul Walker was digitally recreated in the 2015 movie Furious 7; similarly, in 1999, after actor Oliver Reed passed away while filming Gladiator, similar digital recreation technology was used to add Reed's face to shots with a body double.[28] Thus, by applying open-source deep learning models to the once-difficult problem of generating fake audio and video, WaveNet, AudioToObama, and FakeApp transferred deep fake capabilities from movie studios and academics to individuals. The government maintains no export controls or regimes that regulate the spread of movie-making technology, thus demonstrating how regulation could adversely affect other industries or applications.

Lawyers have assessed the legality of generating deep fakes and sharing them on the internet. While some legal experts contend that fake videos are protected by the First Amendment, other lawyers and activists argue that deep fakes may constitute harassment, extortion, or an invasion of privacy.[29] Deep fakes may also face copyright infringement claims depending on the content used as training data for image-to-image translation.[30] Given this unclear legal status and weak ethical grounds, some companies, including Reddit, Gfycat, and Discord, have banned and removed deep fakes.[31] However, each website maintains distinct criteria and mechanisms for removing and identifying verboten content, and detection mechanisms for deep fakes are not standardized across industry. As a result, even with strict terms of use or laws designed to prevent the dissemination of deep fakes, it is effectively impossible to completely remove this content from the internet.

Challenges posed by "viral" videos

One of the key security challenges of deep fakes stems from the possibility of a "viral" (i.e. quickly-spreading) video influencing political processes or causing social unrest. In these cases, waiting hours or days could still allow millions of individuals to view malicious content. As reporting mechanisms and bureaucratic processes may be too slow in these cases, automatic detection technology

28. Cole 2017; Giardina 2017.
29. Chesney and Citron 2018b.
30. Cole 2017.
31. Farokhmanesh 2018.

represents one of the few mechanisms for rectifying this problem. In Chapter 5, we discuss how technical countermeasures could be used to detect or disrupt malicious deep learning applications.

4.5.1 Norms, supply controls, and export controls?

Chapter 3 analyzed how norms and taboos, limiting supply, and export controls were applied to prevent the proliferation of nuclear weapons, biological material, and other dual-use technologies. Could similar mechanisms have prevented deep fake proliferation?

Norms and taboos

Although the Faceswap library warns that the creators maintain a "zero tolerance approach" to malicious uses, the thousands of deep fake videos posted on the internet suggest that this policy has been completely ineffective in controlling the software. Because contributions and downloads are frequently anonymous, it is impossible to impose costs on individual internet users distributed around the world and operating under pseudonyms. As a result, identifying or shaming individuals for contravening usage policies is unfeasible. In the case of nuclear weapons, organizations and governments imposed clear economic, security, and reputational costs on governments or individuals who deviated from norms. Over multiple decades, these norms developed and survived multiple infractions. In contrast, no government or organization exists to create mechanisms for enforcing deep fake norms. Users could provide bottom-up pressure on technology companies to more effectively police fake content; however, even with better detection mechanisms, no catch-all solution exists, and fake content could still be hosted and shared on the internet. This policy contained within the Faceswap library suggests that creating norms is futile for open-source software and could not comprehensively alter individual behavior.

Limiting supply

Limiting supply would also fail to prevent individuals from creating and uploading deep fakes. With tens of millions of photos hosted on the internet and hours of content available for download on YouTube, copious training data exists for creating deep fakes using FakeApp or Faceswap. In the case of biological and nuclear weapons, regulators could ensure compliance through inspections and licensing. Nuclear and biological material can be tracked or detected. In contrast, it is currently

impossible to track the datasets and software libraries used for deep fakes; individuals could even use their own photographs or videos to create deep fakes. Thus, limiting supply may also fail to prevent malicious uses of deep fakes.

Export controls

Altohugh export controls would likely be unsuccessful in limiting deep fake creation or proliferation, tracking or regulating components of deep learning, such as hardware, could yield security benefits. As academic papers, Faceswap, and FakeApp are available for download on the internet, any individual with an internet connection can gain access to the software and data necessary for creating deep fakes; furthermore, deep fakes can be created on relatively inexpensive computing hardware, such as a single graphics processing unit (GPU).[32] In the future, regulators could limit researchers' ability to open-source deep learning models. Alternatively, should future malicious deep learning applications require adversaries to use more powerful computing hardware, such as thousands of GPUs, export controls or regulatory regimes could enable policymakers to track or oversee deep learning use.

4.5.2 Technical countermeasures?

This chapter poses a clear warning to research institutions considering open-sourcing new deep learning models. As described in the preceding subsection, the three methods of limiting dual-use technology proliferation discussed in Chapter 3 - norms, limiting supply, and export controls - would struggle to prevent individuals from creating deep fakes and posting them on the internet. In particular, anonymity and access to open-source projects indicate how internet users can easily create counterfeit videos. Given this ongoing concern for how deep learning is used, the next chapter explores how individuals, companies, and governments could apply technical countermeasures to undermine deep learning models and proactively prevent the creation of deep fakes.

32. Chapter 2 provides an overview of deep learning hardware; a single GPU capable of creating deep fakes could be purchased for under $1000.

Chapter 5

Detection and countermeasures: Countering deep fakes

5.1 Technical countermeasures

In the previous chapter, we studied how three mechanisms for controlling technology - establishing norms and taboos, limiting supply, and regulating exports - may fail to limit deep learning proliferation. In particular, geographic decentralization, anonymity, and open-sourcing pose challenges for policymakers. This chapter analyzes how technical countermeasures could supplement existing mechanisms and mitigate risks from deep learning proliferation. In this thesis, technical countermeasures refer to technology designed to disrupt or delay the training or use of deep learning systems.

5.1.1 Adversarial machine learning and facial recognition

This chapter begins by introducing the nascent field of adversarial machine learning, which studies how machine learning models can train or confuse other machine learning models. Examples of adversarial learning include confusing object recognition models into miscategorizing photos or tricking facial recognition systems into predicting an incorrect identity. In order to contextualize our experiments, we provide a brief history of facial recognition, which was initially used by the FBI and other government agencies in the 1950s to identify suspects.[1] As more data became available

1. "Face Recognition - FBI" 2019.

and sophisticated machine learning techniques were invented, better models, such as Facebook's DeepFace system, were able to match human-level accuracy.[2]

After introducing the mathematics and mechanics for attacking and defending deep learning systems, we apply adversarial techniques to the specific case of facial recognition. Facial recognition systems assign a known identity to a photo of an unknown individual. By attacking and defending a facial recognition model, we demonstrate how technical countermeasures could be used to disrupt other deep learning systems and thus undermine malicious deep learning applications. We performed a series of experiments that analyze how adversarial examples could prevent facial recognition. Adversarial examples are pictures of faces that have been maliciously modified to trick a deep learning system or reduce its confidence. Our experiments significantly reduced model accuracy and confidence, thereby demonstrating how individuals, researchers, or companies could use similar strategies to prevent deep fake creation or undermine other deep learning systems. Furthermore, the technical countermeasures proposed in this chapter also suggest how individuals or companies could prevent uploaded photos from being used to generate deep fakes.

5.2 Technical countermeasures for facial recognition

5.2.1 Facial recognition history: From the FBI to Facebook

Since the 1960s, law enforcement agencies have developed increasingly sophisticated facial recognition software to track and identify suspects.[3] Early algorithms required humans to manually identify the positions of facial landmarks (eyes, ears, nose, and mouth) in photos; law enforcement agents would also quantitatively evaluate photos on other metrics, such as lip thickness. Subsequent methods for facial recognition avoided subjective metrics and instead applied statistical algorithms and emerging machine learning techniques to transform and classify images. In the 1990s, DARPA led efforts to create the FacE REcognition Technology (FERET) program, which directed commercial research and development by offering contracts and standardizing testing on a common database of images.[4] Although the rate of development in commercial facial recognition was significantly accelerated by DARPA's efforts, research datasets remained small; for example,

2. Taigman et al. 2014.
3. "Face Recognition - FBI" 2019.
4. "Face Recognition Technology (FERET) | NIST" 2019.

one 2003 DARPA-led challenge incorporated only 2,413 images. Model accuracy increased during the FERET program but remained low for oblique or unfocused images.[5]

Social media and expanding datasets

As millions of users uploaded billions of photos to social media sites in the 2010s, facial recognition datasets were vastly expanded and used to train more sophisticated models. In particular, Facebook's tagging feature, which allows users to identify the names of individuals pictured in uploaded photos, supplied researchers with copious labeled training samples.[6] In 2014, Facebook's AI lab published a paper describing "DeepFace," a deep learning model for facial recognition that achieved 97.25% accuracy in identifying whether two faces represented the same individual; for comparison, humans were able to identify individuals with an accuracy of 97.53%.[7] Using a nine layer neural network with over 120 million parameters, DeepFace was able to achieve roughly human performance on a dataset of roughly 4,000 individuals across four million photos, which represented the largest facial recognition dataset at the time. Twenty years earlier, DARPA datasets had comprised a few thousand photographs; by 2014, researchers had begun to test models trained on a few million photos. Since DeepFace was announced and demoed at the Conference on Computer Vision and Pattern Recognition (CVPR), deep learning facial recognition models have become widely used and increasingly accessible.

Facial recognition and privacy

Human-level facial recognition, ubiquitous CCTV cameras, and growing internet facial image datasets have prompted privacy-conscious engineers and reporters to warn that accessible facial recognition technology could erode personal freedoms.[8] For example, using cameras and facial recognition, an individual's movement through public spaces could be effortlessly tracked by law enforcement. Given these concerns, regulating facial image data has become a rising concern in recent privacy debates. The definition of biometric data codified in the EU's General Data Protection Regulation (GDPR) explicitly includes "facial images," thus subjecting companies to specific

5. "Face Recognition Technology (FERET) | NIST" 2019.
6. Cohen 2014.
7. Taigman et al. 2014.
8. Schuppe 2018.

compliance procedures and penalties.[9] Companies are also limited by the GDPR in their ability to process facial images, and they may be required to release privacy assessments to consumers.[10] US technology companies also have expressed concerns about facial recognition applications. In a blog post titled "Facial recognition technology: The need for public regulation and corporate responsibility," Brad Smith, Microsoft's current President, raises a series of questions regarding government and private sector use of facial recognition technology.[11] Should law enforcement facial recognition models be subject to minimum accuracy requirements? Should consumers be given warnings that their biometric data (i.e. images of their face) may be collected or shared?

Smith's blog post prompted a slew of media articles on regulating facial recognition; legal and technology experts floated a series of limitations or regulations, including prohibiting facial recognition of children, requiring consent from consumers, and creating legal protections analogous to wiretap orders.[12] In 2017, Washington became the third state - after Illinois and Texas - to regulate commercial biometric identifiers; other states are currently debating similar legislation.[13] Illinois' 2008 Biometric Information Privacy Act remains the only legislation that allows individuals to sue companies for damages related to compromising privacy.[14] As industry leaders and journalists continue to call for regulation, other states and countries may explore similar limitations. However, regardless of this push for biometric privacy, facial recognition models' accessibility indicates how authoritarian states or malicious actors could exploit the technology for nefarious purposes. The remainder of this chapter explores how adversarial examples could be used to undermine facial recognition models and preserve individual privacy.

9. "Art. 4 GDPR - Definitions | General Data Protection Regulation (GDPR)" 2019.
10. Ross 2017.
11. Smith 2018.
12. Brandom 2018.
13. Tumeh 2017.
14. "Top Illinois court says no harm required to sue under biometric data law - Reuters" 2019.

How do deep learning facial recognition models work?

The open-source facial recognition model we tested follows a three-step process for identifying individuals: Preprocessing, encoding, and classifying.

1. **Preprocessing** (face detection, alignment, and cropping): In this first step, a facial recognition system identifies the location of major facial landmarks in a given photo, including the eyes, nose, and mouth. The photo is then rotated, scaled, and cropped to ensure landmarks are placed in a consistent position in the preprocessed photo.

2. **Encoding** (use a neural network to generate a lower dimensional encoding of a face): The preprocessed photo is subsequently used as input data to a neural network. The neural network processes the image through layers of artificial neurons, which perform transformations on the data to convert the input image into a lower dimensional encoding. For example, the original input image could be a 128 by 128 pixel square. If each pixel uses three bytes to represent color, this image contains $128 \cdot 128 \cdot 3 = 49,152$ bytes. The neural network will encode attributes of the face in far less memory, such as 256 bytes.

3. **Classifying** (classify the lower dimensional encoding as a face): In this step, the trained facial recognition system has represented the input image in a lower dimension that requires less memory. Now, another classification model is used to evaluate the probability that a given set of numbers represents a certain individual's face. This step matches the neural network's output to a particular individual.

Implementing facial recognition

Multiple deep learning facial recognition models have been open-sourced and are available for download. One library - called "Face_recognition" - is able to identify one individual's face in six lines of code.[15] Other websites provide tutorials on how to create and use deep learning facial recognition models. In Source Code 5.1, we provide the code required to identify an individual.

15. "ageitgey/face_recognition: The world's simplest facial recognition api for Python and the command line" 2019.

```
1 import face_recognition
2
3 known_image = face_recognition.load_image_file('biden.jpg')
4 unknown_image = face_recognition.load_image_file('unknown.jpg')
5
6 biden_encoding = face_recognition.face_encodings(known_image)[0]
7 unknown_encoding = face_recognition.face_encodings(unknown_image)[0]
8
9 results = face_recognition.compare_faces([biden_encoding], unknown_encoding)
```

Source Code 5.1: Example code used to identify the locations of faces in a photo.[16] The first line loads the facial recognition software library. Lines three and four load known and unknown images of faces; a photo of Joe Biden is in "biden.jpg" and a photo of some unknown individual is in "unknown.jpg." Lines six and seven instruct the facial recognition library to search for faces inside the known and unknown images; line nine compares the mathematical encoding of a face found in "biden.jpg" with a face found in "unknown.jpg."

5.2.2 Undermining facial recognition models

The nascent field of adversarial machine learning has evolved different ways to exploit, confuse, or trick deep learning models. Researchers divide attacks into two categories: Whitebox and blackbox.[17]

- **Whitebox attacks**: In a whitebox attack, an adversary has access to the internal composition and parameter values inside a machine learning model. This allows the attacker to compute mathematical functions that optimize an attack.
- **Blackbox attacks**: In the blackbox scenario, an attacker may only observe the model's output on input data. Although blackbox attacks limit an adversary's knowledge and thus appear more difficult to perform, researchers have demonstrated that it is possible to build a model that approximates the targeted system in order to optimize blackbox attacks.[18]

Attacks on facial recognition models can be performed by modifying input images. By altering pixels on or around an individual's face, an adversary may be able to trick a facial recognition model into identifying a different person. The remainder of this chapter discusses how a deep learning facial recognition model can be disrupted or attacked by adversarial inputs. We focus on a blackbox

17. Li, Yi, and Zhang 2018.
18. Bhagoji et al. 2017.

attack scenario wherein an attacker would not have access to the model's internal parameters or data.

Mathematically optimizing attacks

In the simplest case of an adversarial attack used to undermine a machine learning model, an attacker uses some algorithm to generate perturbations to an input image x, which we assume is correctly classified as an individual with identity y. In a facial recognition system, y is the name of the individual pictured in image x. Using some perturbation function $f(z)$, the attacker constructs a similar image $x' = f(x)$ that looks almost indistinguishable from x but is not classified as y.[19] One attacking scenario can be described by the following optimization problem over the adversarial input x', where $p(y|x')$ is the probability that image x' is classified as individual y:

$$\arg\min_{x'} p(y|x')$$

$$\text{s.t. } ||x' - x|| < \epsilon$$

The first equation encapsulates the attacker's goal of searching for a modified image x' that minimizes the probability of being classified as the correct identity y. The second equation constrains the norm of $x' - x$ to some small amount ϵ, thereby requiring an attacker to generate an image that appears the same to a human viewer but yields an entirely different classification.

Examples of adversarial attacks on images

In a "single-pixel attack," an adversary modifies one pixel of an input image to trigger an incorrect classification.[20] As the attacks can now only change one pixel, $||x' - x||$ will equal only the single modified pixel, thereby guaranteeing that the adversarial image will appear almost indistinguishable from the original. One paper found that an untargeted single-pixel attack algorithm could successfully cause misclassifications for 68.36% and 41.22% of images in two widely used datasets for machine learning, thus indicating how sensitive some deep learning systems and datasets are to minute perturbations. By changing just *one* pixel in an entire image, these attacks could confuse

19. Athalye et al. 2017.
20. Su, Vargas, and Sakurai 2019.

deep learning models.

Other attack algorithms may introduce imperceptible modifications throughout an entire image. The Fast Gradient Sign Method (FGSM), which was first proposed by machine learning researcher Ian Goodfellow in 2014, disturbs an image in the opposite direction of the gradient of a model's loss with respect to a particular input image.[21] Unlike a single-pixel attack, FGSM generates minor changes distributed throughout the entirety of an image. However, as the FGSM algorithm requires knowing or approximating the model's gradient on a particular input, it is considered a whitebox attack.[22]

Academic papers have also demonstrated how adversarial attacks can be effectively applied to confuse real world deep learning systems. In the paper "Synthesizing Robust Adversarial Examples," a group of researchers at MIT 3D-printed physical objects with patterns intended to deceive object recognition models.[23] This attack modified the textures on a three-dimensional digital model; in one example, by perturbing the patterns on a turtle's shell, a deep neural network was fooled into classifying the physical 3D-printed turtle as a rifle. The paper demonstrated how this attack could yield misclassifications from the model across a variety of different orientations. In 2018, a team of AI researchers at Google published a paper "Adversarial Patch" outlining an algorithm for generating a physical patch that, when placed alongside other objects, causes a deep learning model to incorrectly classify images containing both the patch and object.[24] In one experiment, a commonly used object recognition deep neural network called VGG16 was tricked into classifying a banana as a toaster. As the mere presence of this patch was able to compromise the output of the VGG16 object recognition model, the "Adversarial Patch" paper demonstrates how blackbox attacks could be used against real-world systems. For example, an adversary could choose to add adversarial markings to the exterior of a car or to their face to confuse detection systems.

21. Goodfellow, Shlens, and Szegedy 2014.
22. Some papers have explored blackbox FGSM attacks by approximating a model's loss function.
23. Athalye et al. 2017.
24. Brown et al. 2017.

Defending deep learning models

Other research groups have begun to develop defensive techniques for designing and training more robust deep learning systems capable of resisting adversarial attacks. One method called "adversarial training" requires training deep learning systems on a larger dataset that includes both genuine and adversarial examples.[25] For example, instead of simply training a facial recognition system on raw images of individuals' faces, a deep learning model is also trained on perturbed images to anticipate and thwart adversarial attacks. Although some papers have found adversarial training to be effective in certain contexts, others consider it to be a "whack-a-mole" approach to defense as the original model is only exposed to a limited subset of all possible attacks.[26] Later in this chapter, we test adversarial training on a facial recognition model.

Other researchers have proposed methods for detecting and rejecting malicious inputs, or leveraging additional deep learning models to project adversarial examples onto a known range of data.[27] Generally, the expanding field of adversarial machine learning suggests that research will continue to follow a tit-for-tat model as new attacks and defenses are proposed.

5.3 Our attack mechanisms on facial recognition

We designed a set of experiments to demonstrate the feasibility of performing blackbox attacks against a facial recognition deep learning model. This project was partially completed with Michael Karr as a component of our course project for CS 229, Stanford's graduate level machine learning course. After presenting our final research, we were awarded Best Poster in the class of 617 students. We chose to focus on the blackbox scenario as it presents clear parallels to a real-world attack wherein an individual may attempt to fool facial recognition systems without access to the learning model. These attacks could be used by fugitives to evade detection or by spies to confuse enemy systems. More importantly, facial recognition is required to produce deep fakes. Thus, our experiments also suggest methods for how individuals could prevent deep fakes from being created from their uploaded content. For example, before uploading a photo to Facebook, an individual could manually run their photo through our adversarial attack algorithms in order to confuse facial

25. Tramer et al. 2017.
26. Galloway, Tanay, and Taylor 2018.
27. Santhanam and Grnarova 2018.

recognition models.

We tested two attack mechanisms: Perturbing images with random noise, and clustering random noise around facial landmarks.

1. **Attack 1**: In Attack 1, a facial image is perturbed by randomly selecting pixels and perturbing their color. This perturbed image is then aligned and cropped before being used as the input to the facial recognition model. Figure 5.2 provides an example of an image perturbed using Attack 1.

2. **Attack 2**: In Attack 2, two steps are performed: Identifying facial landmarks using a deep learning model and adding random noise around these landmarks. First, our facial landmark detector is used to compute the position of an individual's eyes, nose, mouth, eyebrows, and ears. Then, random pixels around these landmarks are perturbed randomly.[28] This modified image is subsequently used as an input image to our facial recognition model. In Figure 5.3, we provide an example of an input image with random noise clustered around facial landmarks.[29]

We decided to test these attacks on the open-source facial recognition deep learning model released by Cole Murray.[30] Murray's facial recognition model partially relies on the existing object detection "Inception" neural network released by Google;[31] the model also reflects the same preprocessing, encoding, and classifying pipeline discussed earlier in this chapter. We selected this model as it enabled quick prototyping and straightforward modifications to inputs. Furthermore, Google's Inception model is one of the highest accuracy object recognition models.[32] We trained our facial recognition model on the Labeled Faces in the Wild (LFW) dataset, which contains over 13,000

28. We sampled from a 2D Gaussian distribution centered at a given facial landmark to add random noise.

29. Attack 2, which distributes random noise around an individual's facial landmarks, required training and testing another deep learning model to identify facial landmarks. We used the Kaggle "Facial Keypoints Detection Dataset" and the open-source Keras neural network library to design a relatively simple nine-layer neural network to identify facial keypoints. After experimenting with multiple configurations, we saw the best performance (including reasonable training times) from a network that uses one max pooling layer, a flattening layer, two pairs of fully connected into dropout layers, and one final fully connected layer. This model yields an average loss of 2.99 pixels from predicted to actual facial landmark locations. This model was able to generalize well from the Kaggle facial keypoints training dataset to the LFW face images used on our facial recognition deep learning model. Later in this chapter, we discuss some shortcomings of our facial landmark DNN, including its failure to accurately identify mouth features on individuals with beards.

30. Murray 2017.

31. "Going Deeper With Convolutions" 2019.

32. Alemi 2016.

images of over 5,000 individuals.[33] Figure 5.1 provides one sample image in the LFW dataset.

Figure 5.1: One sample image of Prime Minister Tony Blair in the LFW dataset.

Below, we present examples of adversarial images generated using our attack mechanisms.

Figure 5.2: An image perturbed using Attack 1.

Figure 5.3: An image perturbed using Attack 2.

33. "LFW Face Database" 2019.

5.3.1 Results: Attacks lower accuracy and confidence

When trained and tested on the raw images in the LFW dataset, our facial recognition model achieves an accuracy of 94.6%, or it correctly recognizes an individual's identity in 94.6% of all test images.[34] Our model's performance decreases notably on images with random noise added, and most classes demonstrate further decreases in model performance when random perturbations are clustered around facial landmarks. In Table 5.1, we present our facial recognition model's performance on raw and perturbed images for a subset of the individuals in the LFW dataset. In Table 5.2, we present the performance of a modified model on raw and adversarial images; this model was trained on a smaller training dataset that had an equal number of photos of each individual. In Table 5.3, we present the results of performing adversarial training to defend our original facial recognition model. This experiment provided a robustness check for attacks. Table 5.4 reports our model's average confidence on raw images, images perturbed with Attack 1, and images perturbed using Attack 2.

Class name	Raw model	Noisy images	Obscured landmarks
Bill Clinton	0.99	0.75	0.58
George W. Bush	0.98	0.91	0.88
John Negroponte	1.0	0.63	0.63
Hamid Karzai	1.0	0.67	0.50
Tony Blair	0.97	0.69	0.71

Table 5.1: Model performance on raw images, noisy images, and images with obscured landmarks. This model was trained on all images in the LFW dataset and thus had imbalanced class sizes (for example, this model was trained on over 500 images of George W. Bush but only 22 of Hamid Karzai).

34. We used a train-test split of 0.7-0.3, or 70% of our images were used for training and 30% for testing the model. Our facial recognition model was trained on Google Cloud Compute Engine virtual machines.

Class name	Raw model	Noisy images	Obscured landmarks
Bill Clinton	0.88	0.86	0.50
George W. Bush	0.92	0.41	0.15
John Negroponte	0.75	0.38	0.13
Hamid Karzai	1.0	1.0	0.50
Tony Blair	0.93	0.46	0.36

Table 5.2: Model performance on raw images, noisy images, images and images with obscured landmarks. This model was trained on classes limited to 20 images each, or the training dataset consisted of 20 images of each individual.

Class name	Raw model	Noisy images	Obscured landmarks
Bill Clinton	0.63	0.56	0.29
George W. Bush	0.98	0.88	0.60
John Negroponte	1.0	0.50	0.75
Hamid Karzai	0.83	0.58	0.60
Tony Blair	0.88	0.58	0.73

Table 5.3: Adversarial training model performance on raw images, noisy images, and obscured landmarks. This model was trained on both raw and perturbed inputs in order to test whether adversarial training could be used to defend our model from malicious inputs.

Model name	Average confidence
Raw images	0.73
Noisy images	0.65
Obscured landmarks	0.51

Table 5.4: Average confidence of predictions on raw images, noisy images, images and images with obscured landmarks. This model was trained under the same conditions as reported in Table 5.1.

5.3.2 Discussion: Undermining confidence and preventing deep fakes

As reported in Table 5.1, Attack 1 noticeably decreased model accuracy. For some classes, such as Ambassador John Negroponte, the performance gap was dramatic (37%); for others, including President George W. Bush, the trained model did not suffer significantly reduced accuracy (a drop of 7% was observed). One potential explanation for this phenomenon is the use of an imbalanced training dataset: While over 500 photos of George W. Bush were present in our training dataset, our model was trained on only 31 photos of Negroponte. For most classes, including President Bill Clinton and President Hamid Karzai, Attack 2 (clustering noise around facial landmarks) resulted in additional accuracy reductions. However, this decrease in performance was not universal, and accuracy either remained the same or increased slightly for two classes (Prime Minister Tony Blair and Ambassador John Negroponte).

One explanation for Attack 2's varied effectiveness across different classes is our use of different training and testing sets for our facial landmark and facial recognition deep learning models. While the facial recognition model was trained on the Labeled Faces in the Wild (LFW) dataset, the facial landmark model was trained on the Kaggle facial keypoints dataset. Furthermore, as the LFW dataset for facial recognition is roughly twice as large as our facial landmark dataset, the facial landmark model may not have been able to effectively generalize to all images present in the LFW dataset. For example, the facial landmark DNN struggled to predict the position of Hamid Karzai's mouth, perhaps because of his beard. This limitation could explain Attack 2's varied performance across classes. It also demonstrates how deep learning models are sensitive to limitations in their training datasets.

Figure 5.4: The facial landmark model fails to accurately recognize Hamid Karzai's mouth but successfully locates his eyebrows, nose, and eyes.

Training dataset size

Given the results in Table 5.1, we hypothesized that the facial recognition model was significantly more susceptible to attacks on classes that had fewer training images. This hypothesis was motivated by the observation that classes that had many training images available, such as George W. Bush (which has over 500 images), appeared to be less vulnerable to attack. Table 5.2 reports the performance of a model that was trained on classes of equal size: Instead of using a different number of training images for different individuals, this model was trained on the same number of images of John Negroponte, George W. Bush, and all other individuals. This model's performance on unperturbed images is lower than the one tested in Table 5.1; given our use of a smaller training set, this result is unsurprising. For some classes, including George W. Bush, Table 5.2 reveals a significant drop in accuracy for images with obscured facial landmarks. As George W. Bush represents the class with most training images, these results validate our hypothesis and suggests that reducing the number of training images for a given class increases susceptibility to attack.

Robustness check: Adversarial training

Table 5.3 presents our attempt to use adversarial training to defend our deep learning model from adversarial input. Instead of simply training on unperturbed images in the LFW dataset, this model was also trained on images with random noise and images with noise clustered around facial landmarks. Although this model demonstrated higher accuracy in some cases, such obscured facial landmarks for Prime Minister Tony Blair, President Hamid Karzai, and Ambassador John Negroponte, it also resulted in unexpectedly low accuracy in other cases, including its performance on unperturbed images of President Clinton and Prime Minister Blair. Our use of random noise suggests one potential explanation for adversarial training's limited effectiveness: Inconsistency due to randomness across adversarial examples may have made it difficult for our model to highly weight facial features not affected by an attack. This could explain adversarial training's poor performance. In demonstrating how one defense mechanism failed to preserve model performance, these results provide a robustness check for our attack mechanisms.

Dramatically decreasing confidence

In Table 5.4, we report our model's average confidence on predictions for both perturbed and unperturbed images. Our model outputs a final classification for each image based on the highest probability class, where each class indicates a specific individual. Thus, a prediction that an image x contains individual y with 15% confidence may still lead the model to classify x as y if no other classes have a higher confidence. On the dataset of unperturbed LFW images, our model reported a confidence of 73% across all predictions, or it was 73% sure - on average - that each prediction was correct.[35] When Attack 1 was applied to input images, our model's confidence dropped to 65%; when Attack 2 was used, average confidence dropped to 51%. Given the results in Table 5.4, perturbing images randomly and around facial landmarks can significantly reduce prediction confidence. More practically, if facial recognition systems are used by law enforcement agencies for evidence or enforcement, the difference between 73% and 51% confidence could affect the way a given prediction is used. For example, during a law enforcement investigation, 51% accuracy may not represent a threshold for pursuing, arresting, or convicting a suspect. Thus, even if the model continues to identify faces correctly, this reduction in confidence also shows how sensitive the deep learning model is to certain features in an image.

5.4 Implications for technical countermeasures

The empirical results demonstrate how relatively simple adversarial attacks can reduce a deep learning model's accuracy and confidence. As deep learning components remain accessible, technical countermeasures could be used to undermine malicious systems. For example, individuals concerned about invasive deep fakes or anti-privacy facial recognition could selectively modify personal images to prevent their use for deep learning. Before uploading pictures to Facebook, Twitter, or Instagram, individuals could make imperceptible modifications using our attack algorithms in order to confuse a deep learning facial recognition model. Alternatively, websites, such as Facebook or YouTube, could automatically encode adversarial perturbations into all uploaded content. Some researchers have recognized this opportunity: The website "equalAI," which was developed in the MIT Media

35. Note that some of these predictions may have been incorrect.

Lab, offers a tool for overlaying text and applying perturbations to images.[36] The website proposes that individuals run this algorithm to perturb all photos uploaded to the internet in order to better control personal data. Beyond simply preventing facial recognition, these tools could inhibit the creation of deep fakes, which requires identifying facial landmarks and orientations. However, even if these mechanisms are used to perturb new images, millions of photos and videos remain accessible on the internet and could be used as training data. Furthermore, this solution would require different attack mechanisms to be created for different types of data, such as video and audio. In the following chapter, we explore a range of policy implications associated with these experiments, including biometric privacy laws and the equalAI project.

Our experiments also showcase how deep learning has become accessible to internet users. Using cloud computing resources, open-source software, and free datasets, we were able to create a deep learning facial recognition model. This clearly substantiates the proliferation challenges outlined in Chapters 2 and 4. However, relying on disruption techniques and other technical countermeasures to mitigate risk also presents novel challenges. While most existing methods for limiting proliferation are transparent and enforceable among a limited set of actors, countermeasures may fail to disrupt certain malicious deep learning models or unintentionally affect some individuals more than others (for example, in our experiments, images of Ambassador John Negroponte appeared particularly susceptible to random noise perturbations). As research groups around the world develop methods for training more robust models, countermeasures may need to be refined and updated constantly.

36. "equalAIs" 2019.

Chapter 6

Conclusion and policy implications

6.1 Principal conclusions

This thesis examined how individuals and governments could mitigate risks associated with malicious deep learning applications. We proposed three hypotheses: Deep learning's technological components result in a unique proliferation process; existing methods for mitigating dual-use technology proliferation cannot comprehensively limit the spread of deep learning; and technical countermeasures could successfully reduce risk by disrupting or delaying deep learning systems. To support our first hypothesis about deep learning proliferation, we analyzed how large datasets, open-source software libraries, and faster hardware have reduced the technical barriers associated with training deep learning models. We compiled and analyzed a dataset consisting of code from 1,526 published research papers that enabled us to track open-source software and publicly accessible dataset usage; our results clearly establish how a small number of software libraries and large datasets are essential to cutting-edge deep learning research. Furthermore, by analyzing the geographic origin of deep learning papers published at the conference NeurIPS from 1987 to 2018, we concluded that significantly more research institutions distributed throughout the world are conducting deep learning research.

We subsequently examined three existing models for limiting dual-use technology proliferation: Establishing norms and taboos, limiting supply, and applying export controls. In the first case, we studied how norms and taboos effectively changed behavior associated with nuclear weapons development and use. As examples of limiting supply, we analyzed the strengths and shortcomings

of the US Nuclear Regulatory Commission (NRC) and Federal Select Agent Program (SAP), which limit the supply of dual-use technology. In the final case - export controls - we discussed how the US government has created strict regimes, including the International Traffic in Arms Regulations (ITAR) and Export Administration Regulations (EAR), for limiting the export of sensitive dual-use technology. In all three cases, deep learning's anonymity and geographic decentralization pose challenges for enforcing norms, limiting access, and controlling geographic spread.

To support this theory about how existing methods for limiting proliferation would be ineffective in preventing malicious deep learning applications, we investigated how deep fake technology spread from academia to downloadable software in under two years. Following papers published by Stanford and the University of Washington, researchers at NVIDIA open-sourced the models used in their 2017 paper "Unsupervised Image-to-Image Translation Networks."[1] In the months following their release, anonymous developers used the NVIDIA models to release open-source software for generating deep fakes. Other unnamed individuals repackaged this open-source software into a Windows application that could be used to generate deep fakes without any technical experience. Given the national security threats posed by deep fakes, which could "could undermine public trust in recorded images and videos,"[2] we argued that policymakers must pursue a different strategy for limiting dual-use technology proliferation.

Chapter 5 presented a unique technical study on applying technological countermeasures to a facial recognition model in order to undermine its accuracy and confidence in identifying faces. Because our attacks significantly reduced both accuracy and confidence, individuals or companies could use similar methods to prevent facial recognition in images. These attacks could be used to preserve individual privacy and prevent the creation of deep fakes, which requires facial recognition. Similar technical countermeasures could be applied to other deep learning systems as well. In the following sections, we explore the policy implications stemming from these conclusions. In particular, we study how technical and policy mechanisms could be utilized by individuals, companies and universities, governments, and trans-national organizations to counter malicious deep learning systems.

1. Liu, Breuel, and Kautz 2017.
2. "2018-09 ODNI Deep Fakes letter.pdf" 2019.

6.2 External validity and broader implications

A bipartisan consensus has emerged in Washington that deep fakes and deep learning represent serious and growing threats to national security. Senators Mark Warner (D-VA) and Marco Rubio (R-FL) as well as Representatives Adam Schiff (D-CA), Stephanie Murphy (D-FL), and Carlos Curbelo (D-FL) have publicly urged US intelligence agencies investigate this technology and develop countermeasures.[3] Deep fakes could be generated by individuals or foreign governments to spread disinformation before elections, or to cause social or economic unrest by reporting fake new about terrorist attacks or other incidents. Counterfeit videos and audio could be generated to blackmail or extort victims. These concerns do not represent existential threats: Open-source or downloadable technology has made these videos easy to create, even for individuals with no technical skills. Tens of thousands of fake videos are already on the internet. Furthermore, on some online forums, individuals can pay others to create customized deep fakes.[4] This technology's continued unmitigated diffusion constitutes a concerning trend that is already impacting individuals' lives and detrimentally affecting national security. Beyond the specific cases of deep fakes and facial recognition, deep learning's use in fake news, autonomous weapons, and mass surveillance threaten individual privacy, human rights, and information security.

As we explored in Chapter 2, recent trends indicate that data, software, and hardware will only become more accessible, further lowering the technical barriers to developing deep learning systems and raising security and privacy concerns. Facial recognition surveillance systems, which may rely on deep learning models, have been widely deployed in China and other countries. Democratic states, including the United Kingdom, have also tested facial recognition software for law enforcement.[5] On May 14$^{\text{th}}$, 2019, San Francisco banned law enforcement and other government agencies from using facial recognition, citing concerns for potential abuse or privacy infringement.[6] Thus, while some countries have passed more stringent laws for collecting facial data, others have begun to lean on facial recognition for population control and mass surveillance. As a result, deep learning and its applications will continue to be closely intertwined with global security dynamics.

Every year, innovations in deep learning yield more accurate and efficient models capable of

3. "2018-09 ODNI Deep Fakes letter.pdf" 2019; Vincent 2018c.
4. "Mr DeepFakes Forums - Requests" 2019.
5. White 2019.
6. Kate Conger and Kovaleski 2019.

generalizing to new tasks. Our technical experiments and proliferation analysis yield an important lens for considering how policymakers may approach other these new models and AI applications. In particular, our findings on deep learning proliferation and technical countermeasures present clear parallels to other emerging technology. Deep learning has been combined with natural language processing algorithms to generate sophisticated fake text; other models are capable of producing art, imitating satellite images, or faking human fingerprints that trick biometric security mechanisms.[7] Just as deep fake technology spread from academia to individuals, countless other research papers in other area of AI research are open-sourced on GitHub; for example, websites and blogs list hundreds of models capable of translating, transcribing, and generating text.[8] Just as we tracked one such list of research papers for deep learning, policymakers and other researchers could more comprehensively analyze these papers to model dual-use technology proliferation. As these models could also be repurposed for malicious use, this data could bring about important advances in tracking proliferation and developing technical countermeasures.

Yet, as our experiments and case studies reveal, no technical panacea exists for delaying or disrupting deep learning. By tracing deep fakes and confusing a facial recognition model, our work may redefine policymakers' expectations and strategies for controlling similar open-source technology with national security implications. Instead of restrictively limiting development or exports, perhaps policymakers could focus on delaying malicious systems or rendering them unusable: A model with fifty percent accuracy delivers significantly less value than one with over ninety percent accuracy. Furthermore, data on deep learning proliferation could be used to enforce norms of development or software use. Policymakers will need to mix traditional forms of regulation, such as controlling exports and limiting production, with a new focus on technical countermeasures, which could introduce uncertainty or impede malicious actors.

Future research could examine additional cases of deep learning proliferation to develop a robust method for tracking open-source library usage, dataset popularity, and geographic decentralization. These initiatives could provide crucial direction for other researchers to consider how technical countermeasures could bolster national security. In Chapter 4, we tracked the research papers used in open-source or downloadable deep fake applications; additional research could develop a

7. Newman 2018.

8. "keon/awesome-nlp: A curated list of resources dedicated to Natural Language Processing (NLP)" 2019.

systematic method for identifying the research papers used in open source projects. For example, by downloading and comparing the code in open-source projects with the models released in research papers, it may be possible to systematically track deep learning models as they transition from academia to open-source. Alternatively, future projects could focus on extrapolating our work on technical countermeasures to identify common mechanisms that could be used across different types of deep learning models. More specifically, this could include an application or system framework for evaluating the efficacy of different technical countermeasures for disrupting deep learning systems. In the following section on policy implications, we explore additional research opportunities for tracking proliferation and creating technical countermeasures.

6.3 Policy implications

The remainder of this chapter addresses policy implications derived from our three hypotheses: Deep learning proliferates differently than other dual-use technologies, existing methods of limiting proliferation are unlikely to prevent malicious applications of deep learning, and technical countermeasures could be used to undermine or delay malicious applications. We divide policy implications into those relating to technical mechanisms and others analyzing policy tools. Figure 6.1 outlines different technical and policy mechanisms that could limit deep learning proliferation or deter malicious use. In the remainder of this section, we evaluate how policymakers could apply both technical and policy mechanisms to track, delay, or undermine malicious deep learning systems.

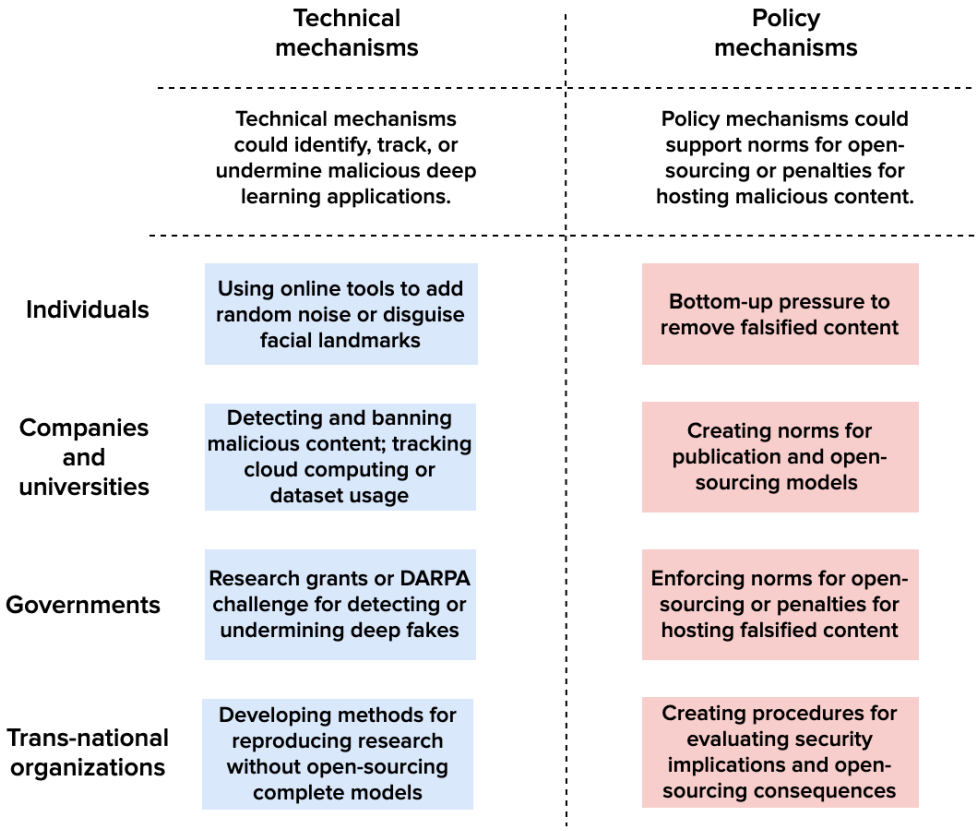

Figure 6.1: Technical and policy mechanisms that individuals, companies and universities, governments, and trans-national organizations could use to limit deep learning proliferation. Policy tools are in red, and technical mechanisms are in blue.

6.3.1 Technical mechanisms for undermining deep learning systems

In Chapter 5, we defined technical countermeasures as technology designed to disrupt or delay deep learning systems; our experiments on an open-source facial recognition model reduced both accuracy and confidence without requiring access to the model's internal parameters or structure. While individuals or companies could apply similar technical countermeasures to prevent the creation of deep fakes, governments could fund or harness technical countermeasures to undermine malicious deep learning applications. Below, we provide greater detail on different ways technical mechanisms could be used to mitigate privacy and security risks from malicious deep learning models.

Individuals could apply technical mechanisms similar to those discussed in Chapter 5 to control how their images and data may be used. For example, before uploading images to social media sites, individuals could apply adversarial perturbations designed to prevent facial recognition. The

MIT Media Lab's EqualAI project already provides a tool designed for this purpose.[9] Furthermore, as social media companies use individuals' personal photos to develop better facial recognition models, these technical mechanisms represent one of the few ways users can assert control over their uploaded data. While Facebook has used uploaded photos to develop better facial recognition software for its own platform, other companies, such as the photo-sharing app Ever, market facial recognition software for enterprise use or government clients. Below, we provide more detail on Ever's use of photos for enterprise facial recognition.[10] Adversarial perturbation mechanisms directly combat this privacy threat.

Ever AI's enterprise facial recognition software

On May 9th, 2019, media outlets reported that the photo storage and sharing application Ever planned to license its facial recognition model - which had been trained on billions of user photos - to private companies and the US government. Using over thirteen billion photos uploaded to their platform, Ever planned to license its "enterprise face recognition" software; the company marketed this software under the entity "Ever AI," explaining:

> "Ever AI's face recognition technology enables smart cities to prepare for the future with start-of-the-art identification and attribution technologies"

This brief excerpt from Ever AI's marketing materials explicitly connects the company's facial recognition product with surveillance programs designed to track individual movement. Although government agencies or companies were not given access to the actual photos used as training data, individuals - likely unwittingly - supplied the requisite training data needed to enable a tool of mass surveillance. As privacy groups and the ACLU criticize Ever, this case clearly validates why individuals may perturb their photos before uploading them to social media platforms.

As a result, politicians, celebrities, and other individuals may choose to run all uploaded photos and videos through a perturbation algorithm designed to deceive facial recognition models. If all official photographs of politicians are adversarially perturbed, using these photos as training data for facial recognition or for creating deep fakes may become more time consuming or computationally

9. "equalAIs" 2019.
10. Khalid 2019.

expensive. Although more sophisticated models may resist these attacks, adversarial perturbations represent one technical countermeasure that could temporarily preserve privacy or make deep fake generation more time consuming, thus aligning with our recommendation that policymakers pursue methods for delaying malicious applications.

Companies could apply similar technical mechanisms on a broader scale to limit malicious deep learning applications. For example, video streaming websites could adversarially perturb all uploaded content to prevent videos from being used to generate deep fakes. Currently, open-source software allows individuals to create deep fakes directly from YouTube videos; one project called "youtube-video-face-swap" promises to "perform a face swap on a youtube video almost automatically."[11] As this online content may be used maliciously, YouTube - or other companies - could add adversarial perturbation mechanisms to prevent this software from automatically generating deep fakes. By increasing technical barriers to creating deep fakes, these systems could serve as an effective deterrent. Although the MIT Media Lab's EqualAIs project is currently intended for individual use, companies could easily add similar technology to their content streaming platforms or allow users to click a button to adversarially perturb uploaded photos or video before releasing them publicly on social media. These technical mechanisms present clear benefits to privacy-conscious individuals and to companies, which can prevent content on their platforms from being used maliciously.

Companies could also apply technical mechanisms to detect counterfeit content. Multiple published papers outline algorithms for detecting deep fakes; for example, one paper monitors individuals' blinking patterns.[12] Although commercial research groups were largely responsible for creating the models that generate deep fakes (as discussed in Chapter 4), no large social media or technology companies have publicly devoted resources to programmatically detecting deep fakes. For example, Facebook, YouTube, and Twitter could develop technology for automatically detecting and removing deep fake videos. This could significantly reduce the national security risks stemming from fast-spreading fake videos. Given these clear benefits to national security, government agencies or policymakers could coordinate and fund detection and removal programs.

11. "DerWaldi/youtube-video-face-swap: The aim of this project ist to perform a face swap on a youtube video almost automatically." 2019.

12. Li et al. 2018.

By funding or requiring the use of technical countermeasures, governments could lead in applying detection systems and technical countermeasures at scale. In 2017 and 2018, DARPA spent $68 million on research projects intended to detect deep fakes.[13] Yet, the agency has been tight-lipped about the capabilities of the technology developed using these funds; thus, it is challenging to assess this program's effectiveness in deterring or undermining malicious deep learning systems. However, our research reveal a series of lessons for how DARPA and other government actors could approach technical countermeasures, including the importance of working with commercial or academic research groups and the risks associated with open-source technology. As commercial and academic researchers were largely responsible for creating the models necessary to generate deep fakes, DARPA could fund projects or competitions for detecting counterfeit content or developing technical countermeasures capable of preventing their creation. DARPA challenges present one model for achieving this goal; the 1990s FERET facial recognition program, Grand Challenge for autonomous vehicles, and robotics challenges for disaster response generated remarkable interest and scientific progress from academic and commercial research groups.[14] Thus, DARPA could consider sponsoring a challenge for developing perturbation mechanisms for preventing deep fakes or facial recognition. Our research also validates the important role secrecy can play in these efforts as well. Ill-conceived open-sourcing or publicization of deep learning technology can yield media scrutiny or facilitate malicious applications, as in the case of NVIDIA's open-sourced models and Adobe VoCo. As a result, even if DARPA decided to sponsor a challenge for undermining deep learning systems, the agency should continue to pursue covert and proprietary technical countermeasures as well.

6.3.2 Policy tools for mitigating deep learning risks

In this section, we examine regulatory and policy mechanisms that could mitigate risks from deep learning proliferation. These policy tools, such as developing mechanisms for evaluating the consequences of open-source models, penalizing companies for hosting or spreading counterfeit content, and protecting personal privacy, could be implemented by companies, governments, and transnational organizations. These policy implications could be implemented alongside the technical

13. Hatmaker 2018.
14. "Face Recognition Technology (FERET) | NIST" 2019.

countermeasures outlined in the previous section.

Companies, universities, and governments should immediately consider methods for systematically evaluating the privacy and national security consequences of deep learning models before they are open-sourced. As demonstrated by NVIDIA's decision to release the models from "Unsupervised Image-to-Image Translation Networks," individuals or research groups may fail to consider the enormous aftereffects that could be triggered by a single research paper and open-source model. As a result, individual companies and universities should develop standards for open-sourcing certain models, sharing datasets, or publishing papers. For example, in February 2019, OpenAI decided not to release a language model due to concerns that it could be used to generate realistic "fake content."[15] Below, we examine OpenAI's reasoning in not releasing their full GPT-2 language model. Other companies and universities could consider similar policies for open-sourcing or releasing limited subsets of research models, which could balance concerns for research reproducibility with the necessity of limiting the second and third order consequences of open-sourcing. Verification procedures could include scrutiny by committees of engineers, policymakers, representatives from academic deep learning conferences, and individuals capable of assessing privacy concerns.

Trans-national organizations could assist in developing and enforcing standards for open-sourcing deep learning models. Reputable deep learning conferences, such as the Conference on Computer Vision and Pattern Recognition (CVPR), the Conference on Neural Information Processing Systems (NeurIPS), or the International Conference on Computer Vision (ICCV), should develop guidelines and advisory committees for evaluating the security externalities of published research that balance the consequences of open-sourcing with the necessity of reproducibility. Other journals in the physical and social sciences have struggled with concerns over research reproducibility, which has prompted some researchers to suggest that papers should release detailed information about all software, data, and processes used to generate published material.[16] Yet, as new deep learning models may pose significant national security threats, these constraints are justified and necessary.

15. "Better Language Models and Their Implications" 2019.
16. Chen et al. 2018.

OpenAI's GPT-2 language model

On February 14th 2019, artificial intelligence research organization OpenAI released a blog post titled "Better Language Models and Their Implications." The article described a new language model called GPT-2, which can perform"reading comprehension, machine translation, question answering, and summarization—all without task-specific training." Using text from eight million web pages, OpenAI researchers trained a model with 1.5 billion parameters, which is roughly ten times more complicated than its predecessor. However, due to concerns about how the model could be used, OpenAI decided against open-sourcing in full. The blog post explained:

> "Due to our concerns about malicious applications of the technology, we are not releasing the trained model. As an experiment in responsible disclosure, we are instead releasing a much smaller model for researchers to experiment with, as well as a technical paper."

After proactively considered the security implications of their new model, which could generate coherent argumentative paragraphs of text, OpenAI decided not to release the full model to the public. Instead, small and medium sized versions of the model were published online; the small version has 117 million parameters and the medium version has 345 million (the full model has 1.5 billion). In an update published in May 2019, the organization also announced that they would share the full model with AI and security partners. OpenAI's decision reflects an evident recognition that open-sourcing can result in malicious uses of technology.

Other companies' updated content policies reflect a similar concern for how their technology or platforms could be used. Reddit, Discord, and other technology companies have banned and removed deep fake videos from their websites.[17] Companies, universities, and research institutions could follow this precedent and more carefully consider the types of content that may be posted online. Although it seems unlikely that code hosting platforms, such as GitHub, BitBucket, and GitLab, which were analyzed in Chapter 2, would remove certain open-source repositories, other social media websites could create more well-defined policies for content removal. For example, YouTube and Twitter could prevent individuals from uploading deep fakes or tutorial videos for

17. Farokhmanesh 2018.

generating them. These tutorial videos significantly reduce the complexity or trial and error required to generate counterfeit content. Furthermore, as technical mechanisms could be used to detect malicious content, these policy mechanisms could operate synergistically with detection and removal software.

Governments could also develop and enforce policies governing data collection, open-sourcing academic papers, and posting fake content. Efforts by individual US states, including Illinois and Ohio, to regulate facial images as biometric data suggest one method for enforcing limitations on data collection; the EU's GDPR legislation, which was analyzed in greater detail in Chapter 5, provides another model for enforcing penalties for illicit data collection.[18] Given these examples, the US federal government could consider enacting nationwide standards similar to statewide programs or the EU's biometric privacy laws.

Policymakers could also develop mechanisms for more vigilantly keeping tabs on deep learning proliferation. Companies could be required or encouraged to track all downloaders of a particular dataset or report users of their cloud computing services. This could enable companies to more aggressively enforce terms of use; for example, training models to produce deep fakes could prompt companies to ban individuals from their computing platforms. However, clear challenges to data collection remain; for example, once an individual is given access to a dataset, they may be able to share it with non-authorized users. Additionally, these techniques may fail to impact well-funded adversaries, who may be able to collect their own data or purchase computing hardware. Yet, as US companies retain large shares of the cloud computing market, tracking cloud computing or dataset users could constitute a first step towards enforcing norms and accountability.

Even though open-sourcing norms may significantly reduce the potential dangers of cutting edge research, policymakers will likely struggle to limit the use of software that is already open-source. Most deep learning software libraries are developed around the world by thousands of contributors; for example, Facebook's popular PyTorch machine learning framework currently has over 1,000 unique contributors and over 17,000 individual code contributions.[19] Collaborators' geographic decentralization suggests that any domestic regulation could be completely evaded by development in other countries, or by contributions from anonymous developers. As a result,

18. Tumeh 2017.

19. "pytorch/pytorch: Tensors and Dynamic neural networks in Python with strong GPU acceleration" 2019.

policymakers should focus on proactively applying technical and policy mechanisms to prevent, delay, or undermine malicious deep learning systems.

6.4 Conclusion

As individuals harness accessible software and hardware to train new models, deep learning's unmitigated diffusion will likely pose additional security threats that warrant inclusion in future threat assessments by the Director of National Intelligence.[20] Deep fakes, which transitioned from academic papers to user-friendly software in under two years, typifies how open-source research can be repurposed and distributed for malicious use. Recent legislation and controversy surrounding facial recognition also indicate how deep learning technology threatens privacy and national security.

By analyzing deep learning proliferation and developing technical countermeasures, this thesis demonstrated how a combination of technical and policy mechanisms could proactively mitigate the national security risks stemming from malicious deep learning systems. In particular, we showed how a combination of existing methods for tracking or limiting technology proliferation could be used in tandem with technological countermeasures. Even though our work suggests policymakers should devote greater attention to technical mechanisms for countering emerging technology, establishing norms of behavior remains essential to reducing national security risks. In the last century, consensus about the dangers posed by nuclear, chemical, and biological weapons yielded collaboration between the government and scientific communities to mitigate the risks associated with proliferation. Today, although deep learning research has become distributed around the world, US organizations, including Facebook, Microsoft, Google, GitHub, Intel, OpenAI, and others, still lead in deep learning research and open-source software development. Thus, even as the number of global institutions conducting deep learning research increases, US policymakers posses a unique ability and responsibility to work with companies to develop norms and requirements for sharing or open-sourcing models and data.

Immediate action is required to track deep learning proliferation, to develop technical countermeasures capable of delaying or disrupting malicious deep learning models, and, most importantly, to establish norms for how new deep learning research is open-sourced. In implementing these

20. Coats 2019.

proposals, policymakers may need to redefine expectations for counter-proliferation. Instead of restrictively limiting use and exports, technical countermeasures could delay or disrupt deep learning systems; our attacks showed how a model with over ninety percent accuracy could be compromised into outputting results no better than a coin flip. Finally, as US intelligence and national security officials scrutinize new deep learning models and research, policymakers must also remember the technology's enormous potential to bring about positive societal impact, from creating art to supporting better healthcare.

Bibliography

"1 Billion Word Language Model Benchmark." 2019. Accessed March 29. http://www.statmt.org/lm-benchmark/.

"[1602.07261] Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning." 2019. Accessed March 4. https://arxiv.org/abs/1602.07261.

"[1606.05250] SQuAD: 100,000+ Questions for Machine Comprehension of Text." 2019. Accessed March 4. https://arxiv.org/abs/1606.05250.

"[1812.04948] A Style-Based Generator Architecture for Generative Adversarial Networks." 2019. Accessed March 4. https://arxiv.org/abs/1812.04948.

"2000 HUB5 English Evaluation Transcripts - Linguistic Data Consortium." 2019. Accessed May 5. https://catalog.ldc.upenn.edu/LDC2002T43.

"2018-09 ODNI Deep Fakes letter.pdf." 2019. Accessed March 30. https://schiff.house.gov/imo/media/doc/2018-09%5C%20ODNI%5C%20Deep%5C%20Fakes%5C%20letter.pdf.

"22 CFR Subchapter M - INTERNATIONAL TRAFFIC IN ARMS REGULATIONS | CFR | US Law | LII / Legal Information Institute." 2019. Accessed March 30. https://www.law.cornell.edu/cfr/text/22/chapter-I/subchapter-M.

"A Beginner's Guide to Generative Adversarial Networks (GANs) | Skymind." 2019. Accessed March 4. https://skymind.ai/wiki/generative-adversarial-network-gan.

"A fast learning algorithm for deep belief nets. - PubMed - NCBI." 2019. Accessed March 4. https://www.ncbi.nlm.nih.gov/pubmed/16764513.

"About ASUS - Facilities & Branches." 2019. Accessed March 30. https://www.asus.com/us/About_ASUS/Facilities-Branches/.

"ageitgey/face_recognition: The world's simplest facial recognition api for Python and the command line." 2019. Accessed April 7. https://github.com/ageitgey/face_recognition.

"AI Research & Development | NVIDIA DGX Systems." 2019. Accessed March 8. https://www.nvidia.com/en-us/data-center/dgx-systems/.

Alemi, Alex. 2016. "Improving inception and image classification in tensorflow." *Google Research Blog* (August 31). Accessed April 29, 2019. Improving%20inception%20and%20image%20classification%20in%20tensorflow.

"Alex Krizhevsky - Google Scholar Citations." 2019. Accessed March 4. https://scholar.google.com/citations?user=xegzhJcAAAAJ&hl=en.

"AlphaGo | DeepMind." 2019. Accessed April 7. https://deepmind.com/research/alphago/.

"An in-depth look at Google's first Tensor Processing Unit (TPU) | Google Cloud Blog." 2019. Accessed March 4. https://cloud.google.com/blog/products/gcp/an-in-depth-look-at-googles-first-tensor-processing-unit-tpu.

"An Introduction to Neural Network Methods for Differential Equations | Neha Yadav | Springer." 2019. Accessed March 4. https://www.springer.com/us/book/9789401798150.

"Announcing $100 million in Series D round funding led by ICONIQ Capital | GitLab." 2019. Accessed April 30. https://about.gitlab.com/2018/09/19/announcing-100m-series-d-funding/.

"Announcing AWS Inferentia: Machine Learning Inference Chip." 2019. Accessed March 8. https://aws.amazon.com/about-aws/whats-new/2018/11/announcing-amazon-inferentia-machine-learning-inference-microchip/.

"Announcing PyTorch 1.0 for both research and production." 2019. Accessed May 5. https://developers.facebook.com/blog/post/2018/05/02/announcing-pytorch-1.0-for-research-production/.

Armerding, Taylor. 2017. "Thieves can steal your voice for authentication." *CSO* (May 16). Accessed March 30, 2019. https://www.csoonline.com/article/3196820/vocal-theft-on-the-horizon.html.

"Art. 4 GDPR - Definitions | General Data Protection Regulation (GDPR)." 2019. Accessed April 7. https://gdpr-info.eu/art-4-gdpr/.

"Artistic Style Transfer with Convolutional Neural Network." 2019. Accessed April 7. https://medium.com/data-science-group-iitr/artistic-style-transfer-with-convolutional-neural-network-7ce2476039fd.

"astorfi/lip-reading-deeplearning: Lip Reading - Cross Audio-Visual Recognition using 3D Architectures." 2019. Accessed April 7. https://github.com/astorfi/lip-reading-deeplearning.

Athalye, Anish, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2017. "Synthesizing robust adversarial examples." *arXiv preprint arXiv:1707.07397.*

"AWS announces new Inferentia machine learning chip | TechCrunch." 2019. Accessed March 4. https://techcrunch.com/2018/11/28/aws-announces-new-inferentia-machine-learning-chip/.

"AWS Cloud Revenue Jumps 45% in Q4, Microsoft Azure Revenue Up 76%." 2019. Accessed March 4. https://www.sdxcentral.com/onlines/news/aws-cloud-revenue-jumps-45-in-q4-microsoft-azure-revenue-up-76/2019/02/.

"AWS vs Azure vs Google Cloud Market Share 2018 Report." 2019. Accessed March 4. https://www.skyhighnetworks.com/cloud-security-blog/microsoft-azure-closes-iaas-adoption-gap-with-amazon-aws/.

Bansal, Aayush, Shugao Ma, Deva Ramanan, and Yaser Sheikh. 2018. "Recycle-gan: Unsupervised video retargeting." In *Proceedings of the European Conference on Computer Vision (ECCV),* 119–135.

"Anfal: Campaign against the Kurds." 2007. *BBC* (June 4). Accessed March 30, 2019. `http://news.bbc.co.uk/2/hi/middle_east/4877364.stm`.

"Adobe Voco 'Photoshop-for-voice' causes concern." 2016. *BBC* (November). Accessed March 30, 2019. `https://www.bbc.com/news/technology-37899902`.

"BBC - iWonder - AI: 15 key moments in the story of artificial intelligence." 2019. Accessed April 7. `https://www.bbc.com/timelines/zq376fr`.

Beaumont, Peter. 2018. "The taboo on chemical weapons has lasted a century - it must be preserved | Peter Beaumont | Opinion | The Guardian." *The Guardian* (April 18). Accessed March 30, 2019. `https://www.theguardian.com/commentisfree/2018/apr/18/chemical-weapons-taboo-syria`.

"Benchmarking Tensorflow Performance and Cost Across Different GPU Options." 2019. Accessed March 28. `https://medium.com/initialized-capital/benchmarking-tensorflow-performance-and-cost-across-different-gpu-options-69bd85fe5d58`.

"benhamner/nips-papers." 2019. Accessed March 8. `https://github.com/benhamner/nips-papers`.

Bernal, Natasha. 2018. "AI can now create life-like human faces from scratch." *The Telegraph* (December 18). Accessed March 4, 2019. `https://www.telegraph.co.uk/technology/2018/12/18/ai-can-now-create-100-per-cent-lifelike-human-faces-scratch/`.

"Better Language Models and Their Implications." 2019. February 14. Accessed March 30, 2019. `https://openai.com/blog/better-language-models/`.

Bhagoji, Arjun Nitin, Warren He, Bo Li, and Dawn Song. 2017. "Exploring the space of black-box attacks on deep neural networks." *arXiv preprint arXiv:1712.09491.*

Borger, Julian, and Peter Beaumont. 2018. "Syria: US, UK and France launch strikes in response to chemical attack." *The Guardian* (April). `https://www.theguardian.com/world/2018/apr/14/syria-air-strikes-us-uk-and-france-launch-attack-on-assad-regime`.

Boylan, Jennifer Finney. 2018. "Will Deep-Fake Technology Destroy Democracy?" *New York Times* (October). `https://www.nytimes.com/2018/10/17/opinion/deep-fake-technology-democracy.html`.

Brandom, Russell. 2018. "How should we regulate facial recognition?" *The Verge* (August 29). Accessed March 15, 2019. `https://www.theverge.com/2018/8/29/17792976/facial-recognition-regulation-rules`.

Brown, Tom B, Dandelion Mane, Aurko Roy, Martin Abadi, and Justin Gilmer. 2017. "Adversarial patch." *arXiv preprint arXiv:1712.09665.*

"BVLC/caffe: Caffe: a fast open framework for deep learning." 2019. Accessed May 5. `https://github.com/BVLC/caffe`.

"Caffe | Deep Learning Framework." 2019. Accessed May 5. http://caffe.berkeleyvision.org/.

Castelvecchi, Davide. 2019. "AI pioneer: 'The dangers of abuse are very real'." *Nature* (April 4). Accessed April 7, 2019. https://www.nature.com/articles/d41586-019-00505-2.

"Cheaper AI for everyone is the promise with Intel and Facebook's new chip - MIT Technology Review." 2019. Accessed March 8. https://www.technologyreview.com/s/612722/cheaper-ai-for-everyone-is-the-promise-with-intel-and-facebooks-new-chip/.

Chen, Xiaoli, Sünje Dallmeier-Tiessen, Robin Dasler, Sebastian Feger, Pamfilos Fokianos, Jose Benito Gonzalez, Harri Hirvonsalo, Dinos Kousidis, Artemis Lavasa, Salvatore Mele, et al. 2018. "Open is not enough." *Nature Physics:* 1.

Chesney, Robert, and Danielle Citron. 2018a. "Deep Fakes: A Looming Crisis for National Security, Democracy, and Privacy?" *Lawfare blog.*

———. 2018b. "Deep Fakes: A Looming Crisis for National Security, Democracy and Privacy?" *Lawfare* (February 21). Accessed March 15, 2019. https://www.lawfareblog.com/deep-fakes-looming-crisis-national-%20security-democracy-and-privacy.

"CIFAR-10 and CIFAR-100 datasets." 2019. Accessed May 5. https://www.cs.toronto.edu/~kriz/cifar.html.

"Cloud costs aren't actually dropping dramatically - Hacker Noon." 2019. Accessed March 4. https://hackernoon.com/cloud-costs-arent-actually-dropping-dramatically-cd94051b021c.

"Cloud Text-to-Speech - Speech Synthesis | Cloud Text-to-Speech API | Google Cloud." 2019. Accessed March 30. https://cloud.google.com/text-to-speech/.

"CNTK on Azure - Cognitive Toolkit - CNTK | Microsoft Docs." 2019. Accessed March 29. https://docs.microsoft.com/en-us/cognitive-toolkit/cntk-on-azure.

"CNTK_2_6_Release_Notes - Cognitive Toolkit - CNTK | Microsoft Docs." 2019. Accessed March 29. https://docs.microsoft.com/en-us/cognitive-toolkit/ReleaseNotes/CNTK_2_6_Release_Notes.

Coats, Daniel R. 2019. "Worldwide Threat Assessment." https://www.dni.gov/files/ODNI/documents/2019-ATA-SFR---SSCI.pdf.

Cohen, David. 2014. "DeepFace: Facebook Uses Artificial Intelligence To Boost Performance Of Facial-Verification Project." *The Next Web* (March 19). Accessed March 15, 2019. https://www.adweek.com/digital/deepface/.

Cole, Samantha. 2017. "AI-Assisted Fake Porn Is Here and We're All Fucked." *Vice* (December 11). Accessed March 30, 2019. https://motherboard.vice.com/en_us/article/gydydm/gal-gadot-fake-ai-porn.

Colombo, Florian, and Wulfram Gerstner. 2018. "BachProp: Learning to Compose Music in Multiple Styles." *arXiv preprint arXiv:1802.05162.*

"Commerce Control List (CCL)." 2019. Accessed March 30. https://www.bis.doc.gov/index.php/regulations/commerce-control-list-ccl.

Commission, US Nuclear Regulatory, et al. 2016. "NRC Vision and Strategy: Safely Achieving Effective and Efficient Non-Light Water Reactor Mission Readiness." *ML16139A812 Draft Rev* 1.

"Companies Reported to Have Sold or Attempted to Sell Libya Gas Centrifuge Components | NTI." 2005. March 1. Accessed March 30, 2019. https://www.nti.org/analysis/articles/companies-sold-libya-gas-centrifuge/.

"Comparison of AI Frameworks | Skymind." 2019. Accessed March 15. https://skymind.ai/wiki/comparison-frameworks-dl4j-tensorflow-pytorch#theano.

"Competitions | Kaggle." 2019. Accessed March 29. https://www.kaggle.com/competitions.

"Cost comparison of deep learning hardware: Google TPUv2 vs Nvidia Tesla V100." 2019. Accessed March 8. https://medium.com/bigdatarepublic/cost-comparison-of-deep-learning-hardware-google-tpuv2-vs-nvidia-tesla-v100-3c63fe56c20f.

"CS231n Convolutional Neural Networks for Visual Recognition." 2019. Accessed March 4. http://cs231n.github.io/convolutional-networks/.

"CUDA Zone | NVIDIA Developer." 2019. Accessed March 15. https://developer.nvidia.com/cuda-zone.

"DARPA is funding new tech that can identify manipulated videos and 'deepfakes' | TechCrunch." 2019. Accessed March 30. https://techcrunch.com/2018/04/30/deepfakes-fake-videos-darpa-sri-international-media-forensics/.

"Data.gov." 2019. Accessed March 29. https://www.data.gov/.

"Datasets | Kaggle." 2019. Accessed March 29. https://www.kaggle.com/datasets.

"Deep Learning 101 - Part 1: History and Background." 2019. Accessed March 4. https://beamandrew.github.io/deeplearning/2017/02/23/deep_learning_101_part1.html.

"Deep Learning and AI frameworks - Azure | Microsoft Docs." 2019. Accessed March 4. https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/dsvm-deep-learning-ai-frameworks.

"Deep learning - User Guide| Alibaba Cloud Documentation Center." 2019. Accessed March 4. https://www.alibabacloud.com/help/doc-detail/75093.html.

"Deep Learning VM Image | Deep Learning VM Image | Google Cloud." 2019. Accessed March 4. https://cloud.google.com/deep-learning-vm/.

"deepfakes guide: Fake App 2 2 Tutorial. installation - YouTube." 2019. Accessed March 30. https://www.youtube.com/watch?v=Lsv38PkLsGU.

"deepfakes/faceswap: Non official project based on original /r/Deepfakes thread." 2019. Accessed March 30. https://github.com/deepfakes/faceswap.

"DeepMind hopes its TensorFlow lib Sonnet is music to ears of AI devs - The Register." 2019. Accessed March 15. https://www.theregister.co.uk/2017/04/07/deepmind_releases_sonnet_its_own_tensorflow_library_for_the_good_of_ai/.

"deepmind/sonnet: TensorFlow-based neural network library." 2019. Accessed May 5. https://github.com/deepmind/sonnet.

"DerWaldi/youtube-video-face-swap: The aim of this project ist to perform a face swap on a youtube video almost automatically." 2019. Accessed May 11. https://github.com/DerWaldi/youtube-video-face-swap.

"Disinformation on Steroids: The Threat of Deep Fakes." 2019. Accessed March 30. https://www.cfr.org/report/deep-fake-disinformation-steroids.

"Document." 2019. Accessed April 30. https://www.sec.gov/Archives/edgar/data/1650372/000165037216000018/a20-f06302016.htm.

"equalAIs." 2019. Accessed April 7. http://equalais.media.mit.edu/.

"Export Administration Regulations (EAR)." 2019. Accessed March 30. https://www.bis.doc.gov/index.php/regulations/export-administration-regulations-ear.

"Face Recognition - FBI." 2019. Accessed March 29. https://www.fbi.gov/file-repository/about-us-cjis-fingerprints_biometrics-biometric-center-of-excellences-face-recognition.pdf/view.

"Face Recognition Technology (FERET) | NIST." 2019. Accessed April 7. https://www.nist.gov/programs-projects/face-recognition-technology-feret.

"Fact Sheet 1 - History." 2019. Accessed March 30. https://www.opcw.org/sites/default/files/documents/Fact_Sheets/English/Fact_Sheet_1_-_History.pdf.

"FakeApp download links and How-To Guide : GifFakes." 2019. Accessed March 26. https://www.reddit.com/r/GifFakes/comments/7xv91x/fakeapp_download_links_and_howto_guide/.

"FakeApp download links and How-To Guide : GifFakes." 2019. Accessed March 30. https://www.reddit.com/r/GifFakes/comments/7xv91x/fakeapp_download_links_and_howto_guide/.

Farokhmanesh, Megan. 2018. "Deepfakes are disappearing from parts of the web, but they're not going away." *The Verge* (February 9). Accessed March 30, 2019. https://www.theverge.com/2018/2/9/16986602/deepfakes-banned-reddit-ai-faceswap-porn.

"Fast Track Action Committee Report : Recommendations on the Select Agent Regulations Based on Broad Stakeholder Engagment - October 2015." 2019. Accessed March 30. https://www.phe.gov/s3/Documents/ftac-sar.pdf.

"fast.ai Datasets | fast.ai course v3." 2019. Accessed March 8. https://course.fast.ai/datasets.

"Federal Select Agent Program - About Us." 2019. Accessed March 30. https://www.selectagents.gov/about.html.

Ferris, Patrick. 2018. "An introduction to explainable AI, and why we need it." *Medium* (August 27). Accessed March 28, 2019. https://medium.freecodecamp.org/an-introduction-to-explainable-ai-and-why-we-need-it-a326417dd000.

Fitzgerald, Gerard J. 2008. "Chemical warfare and medical response during World War I." *American journal of public health* 98 (4): 611–625.

"Fueling the Gold Rush: The Greatest Public Datasets for AI." 2019. Accessed March 8. https://medium.com/startup-grind/fueling-the-ai-gold-rush-7ae438505bc2.

Galloway, Angus, Thomas Tanay, and Graham W Taylor. 2018. "Adversarial training versus weight decay." *arXiv preprint arXiv:1804.03308.*

"GeForce GTX 1080 Ti Graphics Cards | NVIDIA GeForce." 2019. Accessed March 28. https://www.nvidia.com/en-us/geforce/products/10series/geforce-gtx-1080-ti/.

"Generative Models." 2019. Accessed March 4. https://blog.openai.com/generative-models/.

"Get Started with Deep Learning Using the AWS Deep Learning AMI | AWS Machine Learning Blog." 2019. Accessed March 4. https://aws.amazon.com/blogs/machine-learning/get-started-with-deep-learning-using-the-aws-deep-learning-ami/.

"Get Started with Oracle Machine Learning." 2019. Accessed March 4. https://docs.oracle.com/en/cloud/paas/autonomous-data-warehouse-cloud/omlug/get-started-oracle-machine-learning.html#GUID-2AEC56A4-E751-48A3-AAA0-0659EDD639BA.

"Getting started tutorial." 2019. Accessed March 4. https://console.bluemix.net/docs/services/PredictiveModeling/index.html#WMLgettingstarted.

Giardina, Carolyn. 2017. "'Furious 7' and How Peter Jackson's Weta Digitally Completed Paul Walker | Hollywood Reporter." *The Hollywood Reporter* (March 25). Accessed March 30, 2019. https://www.hollywoodreporter.com/behind-screen/furious-7-how-peter-jacksons-784157.

"GitHub vs. Bitbucket vs. GitLab vs. Coding - flow.ci - Medium." 2019. Accessed March 29. https://medium.com/flow-ci/github-vs-bitbucket-vs-gitlab-vs-coding-7cf2b43888a1.

"Going Deeper With Convolutions." 2019. Accessed March 4. http://www.cs.unc.edu/~wliu/papers/GoogLeNet.pdf.

Gokani, Jaykishen. 2019. "The Evolution of Banking: AI | MS&E 238 Blog." Accessed March 28. https://mse238blog.stanford.edu/2017/08/jgokani/the-evolution-of-banking-ai/.

Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. 2014. "Explaining and harnessing adversarial examples." *arXiv preprint arXiv:1412.6572.*

Goodin, Dan. 2019. "Widely used open source software contained bitcoin-stealing backdoor | Ars Technica." Accessed March 29. https://arstechnica.com/information-technology/2018/11/hacker-backdoors-widely-used-open-source-software-to-steal-bitcoin/.

"google/filament: Filament is a real-time physically based rendering engine for Android, iOS, Windows, Linux, macOS and WASM/WebGL." 2019. Accessed March 29. https://github.com/google/filament.

"google/uis-rnn: This is the library for the Unbounded Interleaved-State Recurrent Neural Network (UIS-RNN) algorithm, corresponding to the paper Fully Supervised Speaker Diarization." 2019. Accessed March 29. https://github.com/google/uis-rnn.

Goyal, Priya, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. "Accurate, large minibatch sgd: Training imagenet in 1 hour." *arXiv preprint arXiv:1706.02677*.

"GPU computing: Accelerating the deep learning curve | ZDNet." 2019. Accessed March 15. https://www.zdnet.com/article/gpu-computing-accelerating-the-deep-learning-curve/.

Grip, Lina, and John Hart. 2009. "The use of chemical weapons in the 1935–36 Italo-Ethiopian War." *SIPRI Arms Control and Non-proliferation Programmme:* 8.

"Hands-on Labs Image Recognition - Cognitive Toolkit - CNTK | Microsoft Docs." 2019. Accessed March 4. https://docs.microsoft.com/en-us/cognitive-toolkit/Hands-On-Labs-Image-Recognition.

Hatmaker, Taylor. 2017. "ITAR Compliance Requirements You Need To Know." *The Next Web* (July 5). Accessed March 15, 2019. https://www.ftptoday.com/blog/itar-compliance-requirements-you-need-to-know.

———. 2018. "DARPA is funding new tech that can identify manipulated videos and 'deepfakes'." *The Next Web* (April). Accessed March 15, 2019. https://techcrunch.com/2018/04/30/deepfakes-fake-videos-darpa-sri-international-media-forensics/.

"History of Neural Networks." 2019. Accessed March 4. http://www.psych.utoronto.ca/users/reingold/courses/ai/cache/neural4.html.

"Home - Keras Documentation." 2019. Accessed March 29. https://keras.io/.

House, Matthew. 2013. "The Select Agent Program and dual use research of concern: their effects on the regulatory environment of pandemic influenza studies."

"How many cores in a standard gpu? - Quora." 2019. Accessed March 29. https://www.quora.com/How-many-cores-in-a-standard-gpu.

"How to get Images from ImageNet with Python in Google Colaboratory." 2019. Accessed March 28. https://medium.com/coinmonks/how-to-get-images-from-imagenet-with-python-in-google-colaboratory-aeef5c1c45e5.

"How To Install FakeApp - Alan Zucconi." 2019. Accessed March 30. https://www.alanzucconi.com/2018/03/14/how-to-install-fakeapp/.

"IBM100 - Deep Blue." 2019. Accessed April 7. https://www.ibm.com/ibm/history/ibm100/us/en/icons/deepblue/.

"Image Classification using Flowers dataset | Cloud ML Engine for TensorFlow | Google Cloud." 2019. Accessed March 29. https://cloud.google.com/ml-engine/docs/tensorflow/flowers-tutorial.

"ImageNet." 2019. Accessed March 29. http://www.image-net.org/about-stats.

"ImageNet." 2019. Accessed May 5. http://image-net.org/about-overview.

"ImageNet Classification with Deep Convolutional Neural Networks." 2019. Accessed March 4. https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

"ImageNet: the data that spawned the current AI boom - Quartz." 2019. Accessed March 4. https://qz.com/1034972/the-data-that-changed-the-direction-of-ai-research-and-possibly-the-world/.

"Intel details Nervana, a neural network chip for inference-based workloads (Updated) | VentureBeat." 2019. Accessed March 8. https://venturebeat.com/2019/01/07/intel-announces-nervana-a-neural-network-chip-for-inference-based-workloads/.

"Introducing Activation Atlases." 2019. Accessed March 29. https://openai.com/blog/introducing-activation-atlases/.

"Introduction to the Python Deep Learning Library Theano." 2019. Accessed May 5. https://machinelearningmastery.com/introduction-python-deep-learning-library-theano/.

Jones, Gary. 1998. *Nuclear Regulatory Commission Preventing Problem Plants Requires More Effective Action by NRC.* Technical report. U.S. Government Accountability Office.

Kate Conger, Richard Fausset, and Serge F. Kovaleski. 2019. "San Francisco Bans Facial Recognition Technology." *The New York Times* (May 14). Accessed May 14, 2019. https://www.nytimes.com/2019/05/14/us/facial-recognition-ban-san-francisco.html.

"keon/awesome-nlp: A curated list of resources dedicated to Natural Language Processing (NLP)." 2019. Accessed May 14. https://github.com/keon/awesome-nlp.

Khalid, Amrita. 2019. "Ever quietly trained facial recognition AI using its photo storage app." *Engadget* (May 9). Accessed May 11, 2019. https://www.engadget.com/2019/05/09/ever-photo-storage-app-facial-recognition/.

Knight, Will. 2018. "An AI-driven robot hand spent a hundred years teaching itself to rotate a cube." July 30. Accessed April 7, 2019. https://www.technologyreview.com/s/611724/artificial-intelligence-driven-robot-hand-spends-a-hundred-years-teaching-itself-to-rotate/.

"Google's self-driving-car project becomes a separate company: Waymo - Los Angeles Times." 2016. *LA Times* (December 13). Accessed April 7, 2019. https://www.latimes.com/business/autos/la-fi-hy-google-waymo-self-driving-20161213-story.html.

Lee, Dave. 2018a. "Deepfakes are disappearing from parts of the web, but they're not going away." *The Verge* (February). https://www.theverge.com/2018/2/9/16986602/deepfakes-banned-reddit-ai-faceswap-porn.

———. 2018b. "Deepfakes porn has serious consequences." *BBC* (February). https://www.bbc.com/news/technology-42912529.

Lee, Timothy B. 2018. "How computers got shockingly good at recognizing images." *Ars Technica* (December 18). Accessed March 15, 2019. https://arstechnica.com/science/2018/12/how-computers-got-shockingly-good-at-recognizing-images/3/.

"LFW Face Database." 2019. Accessed April 7. http://vis-www.cs.umass.edu/lfw/.

Li, Pengcheng, Jinfeng Yi, and Lijun Zhang. 2018. "Query-Efficient Black-Box Attack by Active Learning." *arXiv preprint arXiv:1809.04913.*

Li, Yuezun, Ming-Ching Chang, Hany Farid, and Siwei Lyu. 2018. "In ictu oculi: Exposing ai generated fake face videos by detecting eye blinking." *arXiv preprint arXiv:1806.02877.*

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. "Microsoft coco: Common objects in context." In *European conference on computer vision,* 740–755. Springer.

Liu, Ming-Yu, Thomas Breuel, and Jan Kautz. 2017. "Unsupervised image-to-image translation networks." In *Advances in Neural Information Processing Systems,* 700–708.

"LSUN." 2019. Accessed May 5. https://www.yf.io/p/lsun.

Mak, Tim. 2018. "Can You Believe Your Own Ears? With New 'Fake News' Tech, Not Necessarily." *NPR* (April 4). Accessed March 30, 2019. https://www.npr.org/2018/04/04/599126774/can-you-believe-your-own-ears-with-new-fake-news-tech-not-necessarily.

Martinsson, Johanna. 2011. *Global Norms: Creation, Diffusion, and Limits.*

"NIPS Accepted Papers Stats - Machine Learning in Practice." 2017. December 5. Accessed March 8, 2019. https://medium.com/machine-learning-in-practice/nips-accepted-papers-stats-26f124843aa0.

"8 Best Deep Learning Frameworks for Data Science enthusiasts." 2018. *Medium* (April 5). Accessed March 15, 2019. https://medium.com/the-mission/8-best-deep-learning-frameworks-for-data-science-enthusiasts-d72714157761.

Metz, Cade. 2015. "Google Just Open Sourced TensorFlow, Its Artificial Intelligence Engine." *Wired* (November 9). Accessed March 4, 2019. https://www.wired.com/2015/11/google-open-sources-its-artificial-intelligence-engine/.

"Microsoft Azure Revenues Climb 76% | Light Reading." 2019. Accessed March 4. https://www.lightreading.com/enterprise-cloud/infrastructure-and-platform/microsoft-azure-revenues-climb-76-/d/d-id/749163?_mc=RSS_LR_EDT.

"mingyuliutw/UNIT: Unsupervised Image-to-Image Translation." 2019. Accessed March 30. https://github.com/mingyuliutw/UNIT.

"MNIST handwritten digit database, Yann LeCun, Corinna Cortes and Chris Burges." 2019. Accessed May 5. http://yann.lecun.com/exdb/mnist/.

"Mr DeepFakes Forums - Requests." 2019. Accessed May 11. https://mrdeepfakes.com/forums/forum-requests.

Murray, Cole. 2017. "Building a Facial Recognition Pipeline with Deep Learning in Tensorflow." *Hackernoon* (August 7). Accessed April 7, 2019. https://hackernoon.com/building-a-facial-recognition-pipeline-with-deep-learning-in-tensorflow-66e7645015b8?gi=5b7ce7a01731.

"MXNet: A Scalable Deep Learning Framework." 2019. Accessed May 5. https://mxnet.apache.org/.

Newman, Lily Hay. 2018. "Machine Learning Can Create Fake 'Master Key' Fingerprints." *Wired* (November 17). Accessed May 14, 2019. https://www.wired.com/story/deepmasterprints-fake-fingerprints-machine-learning/.

"NRC: Backgrounder on Nuclear Power Plant Licensing Process." 2019. Accessed March 30. https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/licensing-process-fs.html.

"NRC: Backgrounder on the Three Mile Island Accident." 2019. Accessed March 30. https://www.nrc.gov/reading-rm/doc-collections/fact-sheets/3mile-isle.html.

"NRC: Medical, Industrial, & Academic Uses of Nuclear Materials." 2019. Accessed March 30. https://www.nrc.gov/materials/medical.html.

"NRC: Source Material." 2019. Accessed March 30. https://www.nrc.gov/materials/srcmaterial.html.

"NRC Vision And Strategy For Licensing Advanced Reactors Needs Improvement." 2019. Accessed March 30. https://www.forbes.com/sites/rodadams/2017/01/09/nrc-vision-and-strategy-for-licensing-advanced-reactors-needs-improvement/.

"Nuclear weapons timeline | ICAN." 2019. Accessed March 30. http://www.icanw.org/the-facts/the-nuclear-age/.

"NUREG-1542, Vol.21, Suppl 1, "Fiscal Year 2015, Summary of Performance and Financial Information."" 2019. Accessed March 30. https://www.nrc.gov/docs/ML1604/ML16047A356.pdf.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. 2016. "Wavenet: A generative model for raw audio." *arXiv preprint arXiv:1609.03499.*

"Open sourcing Sonnet - a new library for constructing neural networks | DeepMind." 2019. Accessed March 15. https://deepmind.com/blog/open-sourcing-sonnet/.

"openai/gpt-2: Code for the paper "Language Models are Unsupervised Multitask Learners"." 2019. Accessed March 30. https://github.com/openai/gpt-2.

"openslr.org." 2019. Accessed March 29. http://www.openslr.org/12/.

Oversight of Security at Commercial Nuclear Power Plants Needs to Be Strengthened. 1998. Technical report. U.S. Government Accountability Office.

"Overview of Neuron Structure and Function - Molecular Cell Biology - NCBI Bookshelf." 2019. Accessed March 4. https://www.ncbi.nlm.nih.gov/books/NBK21535/.

Patel, Prince. 2018. "Why Python is the most popular language used for Machine Learning." March 8. Accessed March 28, 2019. https://medium.com/@UdacityINDIA/why-use-python-for-machine-learning-e4b0b4457a77.

"Perceptron Reading Material." 2019. Accessed March 4. http://www.cs.cmu.edu/~10701/slides/Perceptron_Reading_Material.pdf.

Price, Richard. 1995a. "A genealogy of the chemical weapons taboo." *International Organization* 49 (1): 73–103.

———. 1995b. "A Genealogy of the Chemical Weapons Taboo." *International Organization* 49 (1): 73–103. ISSN: 00208183, 15315088. http://www.jstor.org/stable/2706867.

Prohibition of Chemical Weapons+ 31 70 416 3300 http://www. opcw. org, Organisation for the. 2015. "Protocol for the prohibition of the use in war of asphyxiating, poisonous, or other gases, and of bacteriological methods of warfare, 1925 (Geneva Protocol of 1925)." *OPCW: The Legal Texts:* 737–737.

"Public Health Security and Bioterrorism Preparedness and Response Act of 2002 | Department of Energy." 2019. Accessed March 30. https://www.energy.gov/ehss/downloads/public-health-security-and-bioterrorism-preparedness-and-response-act-2002.

"PyTorch Releases Major Update, Now Officially Supports Windows." 2019. Accessed March 4. https://medium.com/syncedreview/pytorch-releases-major-update-now-officially-supports-windows-2426c9f29d2d.

"pytorch/pytorch: Tensors and Dynamic neural networks in Python with strong GPU acceleration." 2019. Accessed April 19. https://github.com/pytorch/pytorch.

Rajpurkar, Pranav, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. "Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning." *arXiv preprint arXiv:1711.05225.*

"Report of the Federal Experts Security Advisory Panel - December 2014." 2019. Accessed March 30. https://www.phe.gov/s3/Documents/fesap.pdf.

Robertson, Adi. 2019. "I'm using AI to face-swap Elon Musk and Jeff Bezos, and I'm really bad at it." *The Verge.* Accessed March 30. https://www.theverge.com/2018/2/11/16992986/fakeapp-deepfakes-ai-face-swapping.

Rogers, Everett M. 1962. *Diffusion of innovations.* Simon / Schuster.

Rosenblith, Walter A. 1961. "On some social consequences of scientific and technological change." *Daedalus* 90 (3): 498–513.

Ross, Danny. 2017. "Processing biometric data? Be careful, under the GDPR." *IAPP* (August 31). Accessed March 15, 2019. https://iapp.org/news/a/processing-biometric-data-be-careful-under-the-gdpr.

Sagan, Scott D. 1997. "Why do states build nuclear weapons? Three models in search of a bomb." *International security* 21 (3): 54–86.

Santhanam, Gokula Krishnan, and Paulina Grnarova. 2018. "Defending against adversarial attacks by leveraging an entire GAN." *arXiv preprint arXiv:1805.10652.*

Scarpino, Matthew. 2018. *Tensorflow for Dummies.* John Wiley & Sons.

Schuppe, Jon. 2018. "Facial recognition gives police a powerful new tracking tool. It's also raising alarms." *The Next Web* (July 30). Accessed March 15, 2019. https://www.nbcnews.com/news/us-news/facial-recognition-gives-police-powerful-new-tracking-tool-it-s-n894936.

Schwartz, Oscar. 2018. "You thought fake news was bad? Deep fakes are where truth goes to die." *The Guardian* (November). https://www.theguardian.com/technology/2018/nov/12/deep-fakes-fake-news-truth.

"Select agents & toxins FAQ." 2019. Accessed March 30. https://www.selectagents.gov/faq-general.html.

"shaoanlu/faceswap-GAN: A denoising autoencoder + adversarial losses and attention mechanisms for face swapping." 2019. Accessed March 4. https://github.com/shaoanlu/faceswap-GAN.

Simonite, Tom. 2014. "Facebook Creates Software That Matches Faces Almost as Well as You Do." *MIT Technology Review* (March 7). Accessed April 7, 2019. https://www.technologyreview.com/s/525586/facebook-creates-software-that-matches-faces-almost-as-well-as-you-do/.

Smith, Brad. 2018. "Facial Recognition Technology: The Need for Public Regulation and Corporate Responsibility." *Microsoft, July* 13 (July 13). Accessed April 29, 2019. https://blogs.microsoft.com/on-the-issues/2018/07/13/facial-recognition-technology-the-need-for-public-regulation-and-corporate-responsibility/.

"Software Practice and Experience | RG Impact Rankings 2018 and 2019." 2019. Accessed March 15. https://www.researchgate.net/journal/1097-024X_Software_Practice_and_Experience.

"Stanford Machine Learning Group." 2019. Accessed March 29. https://github.com/stanfordmlgroup.

"Stanford NLP." 2019. Accessed March 29. https://github.com/stanfordnlp.

"Stanford University CS231n: Convolutional Neural Networks for Visual Recognition." 2019. Accessed April 7. http://cs231n.stanford.edu/.

Su, Jiawei, Danilo Vasconcellos Vargas, and Kouichi Sakurai. 2019. "One pixel attack for fooling deep neural networks." *IEEE Transactions on Evolutionary Computation.*

Suwajanakorn, Supasorn, Steven M Seitz, and Ira Kemelmacher-Shlizerman. 2017. "Synthesizing obama: learning lip sync from audio." *ACM Transactions on Graphics (TOG)* 36 (4): 95.

"Swift Documentation Quick Guide." 2019. Accessed March 8. https://useyourloaf.com/blog/swift-documentation-quick-guide/.

Taigman, Yaniv, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. "Deepface: Closing the gap to human-level performance in face verification." In *Proceedings of the IEEE conference on computer vision and pattern recognition,* 1701–1708.

Tannenwald, Nina. 2005. "Stigmatizing the bomb: Origins of the nuclear taboo." *International Security* 29 (4): 5–49.

Tannenwald, Nina. 2007. *The nuclear taboo: The United States and the non-use of nuclear weapons since 1945.* Vol. 87. Cambridge University Press.

Temples, James R. 1982. "The Nuclear Regulatory Commission and the Politics of Regulatory Reform: Since Three Mile Island." *Public Administration Review:* 355–362.

"TensorFlow – opensource.google.com." 2019. Accessed May 5. `https://opensource.google.com/projects/tensorflow`.

"TensorFlow performance test: CPU VS GPU - Andriy Lazorenko - Medium." n.d. `https://medium.com/@andriylazorenko/tensorflow-performance-test-cpu-vs-gpu-79fcd39170c`.

"TensorFlow performance test: CPU VS GPU - Andriy Lazorenko - Medium." 2019. Accessed March 4. `https://medium.com/@andriylazorenko/tensorflow-performance-test-cpu-vs-gpu-79fcd39170c`.

"tensorflow/tensorflow: An Open Source Machine Learning Framework for Everyone." 2019. Accessed March 4. `https://github.com/tensorflow/tensorflow`.

"The major advancements in Deep Learning in 2018 | Tryolabs Blog." 2019. Accessed March 4. `https://tryolabs.com/blog/2018/12/19/major-advancements-deep-learning-2018/`.

"The Microsoft Cognitive Toolkit - Cognitive Toolkit - CNTK | Microsoft Docs." 2019. Accessed May 5. `https://docs.microsoft.com/en-us/cognitive-toolkit/`.

"The Stanford Question Answering Dataset." 2019. Accessed May 5. `https://rajpurkar.github.io/SQuAD-explorer/`.

"The State of the Octoverse | The State of the Octoverse reflects on 2018 so far, teamwork across time zones, and 1.1 billion contributions." 2019. Accessed March 29. `https://octoverse.github.com/`.

Thies, Justus, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. "Face2face: Real-time face capture and reenactment of rgb videos." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 2387–2395.

"Timeline of Syrian Chemical Weapons Activity, 2012-2019 | Arms Control Association." 2019. Accessed March 30. `https://www.armscontrol.org/factsheets/Timeline-of-Syrian-Chemical-Weapons-Activity`.

"Top Illinois court says no harm required to sue under biometric data law - Reuters." 2019. Accessed April 7. `https://www.reuters.com/article/employment-illinois/top-illinois-court-says-no-harm-required-to-sue-under-biometric-data-law-idUSL1N1ZP1O5`.

"Training ResNet on Cloud TPU | Cloud TPU | Google Cloud." 2019. Accessed March 4. `https://cloud.google.com/tpu/docs/tutorials/resnet`.

Tramer, Florian, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. 2017. "Ensemble adversarial training: Attacks and defenses." *arXiv preprint arXiv:1705.07204.*

"Treaty on the Non-Proliferation of Nuclear Weapons (NPT) - UNODA." 2019. Accessed March 30. `https://www.un.org/disarmament/wmd/nuclear/npt/text`.

Tribe, Laurence H. 1973. *Channeling technology through law.* Bracton Press.

Tumeh, Lara. 2017. "Washington's New Biometric Privacy Statute and How It Compares to Illinois and Texas Law." *JDSupra* (October 20). Accessed April 7, 2019. https://www.jdsupra.com/legalnews/washington-s-new-biometric-privacy-70894/.

Tushe, Nadine. 2011. "US export controls: Do they undermine the competitiveness of US companies in the transatlantic defense market?" *Public Contract Law Journal:* 57–73.

Villasenor, John. 2019. "Artificial intelligence, deepfakes, and the uncertain future of truth." *Brookings* (February). https://www.brookings.edu/blog/techtank/2019/02/14/artificial-intelligence-deepfakes-and-the-uncertain-future-of-truth/.

Vincent, James. 2016. "Baidu follows US tech giants and open sources its deep learning tools." *The Verge* (September 1). Accessed May 5, 2019. https://www.theverge.com/2016/9/1/12725804/baidu-machine-learning-open-source-paddle.

———. 2018a. "Artificial intelligence is going to supercharge surveillance." *The Verge* (January 23). Accessed April 7, 2019. https://www.theverge.com/2018/1/23/16907238/artificial-intelligence-surveillance-cameras-security.

———. 2018b. "The people making fake AI porn have been temporarily distracted by Nicolas Cage." *The Verge* (January 29). Accessed March 30, 2019. https://www.theverge.com/tldr/2018/1/29/16944474/fake-ai-porn-nicolas-cage-reddit.

———. 2018c. "US lawmakers say AI deepfakes 'have the potential to disrupt every facet of our society'." *The Verge* (September 14). Accessed May 11, 2019. https://www.theverge.com/2018/9/14/17859188/ai-deepfakes-national-security-threat-lawmakers-letter-intelligence-community.

———. 2019a. "How three French students used borrowed code to put the first AI portrait in Christie's." *The Verge.* Accessed March 4. https://www.theverge.com/2018/10/23/18013190/ai-art-portrait-auction-christies-belamy-obvious-robbie-barrat-gans.

———. 2019b. "New AI research makes it easier to create fake footage of someone speaking." Accessed March 30. https://www.theverge.com/2017/7/12/15957844/ai-fake-video-audio-speech-obama.

———. 2019c. "Why we need a better definition of 'deepfake'." *The Verge.* Accessed March 30. https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news.

"Vision API - Image Content Analysis | Cloud Vision API | Google Cloud." 2019. Accessed April 7. https://cloud.google.com/vision/.

"#VoCo. Adobe MAX 2016 (Sneak Peeks) | Adobe Creative Cloud - YouTube." 2019. Accessed March 30. https://www.youtube.com/watch?v=I3l4XLZ59iw.

"What is Machine Learning? - Introduction | Coursera." 2019. Accessed March 29. https://www.coursera.org/lecture/machine-learning/what-is-machine-learning-Ujm7v.

"What is the difference between CUDA core and CPU core? - Stack Overflow." 2019. Accessed March 29. https://stackoverflow.com/questions/20976556/what-is-the-difference-between-cuda-core-and-cpu-core.

"What Killed the Curse of Dimensionality? - Hacker Noon." 2019. Accessed March 4. https://hackernoon.com/what-killed-the-curse-of-dimensionality-8dbfad265bbe.

White, Geoff. 2019. "Use of facial recognition tech 'dangerously irresponsible'." *BBC News* (May 13). Accessed May 11, 2019. https://www.bbc.com/news/technology-48222017.

"Who's Ahead in AI Research? Insights from NIPS, Most Prestigious AI Conference." 2019. Accessed May 5. https://medium.com/@chuvpilo/whos-ahead-in-ai-research-insights-from-nips-most-prestigious-ai-conference-df2c361236f6.

Zhang, Richard, Phillip Isola, and Alexei A Efros. 2016. "Colorful image colorization." In *European conference on computer vision,* 649–666. Springer.

"zziz/pwc: Papers with code. Sorted by stars. Updated weekly." 2019. Accessed March 29. https://github.com/zziz/pwc.