



NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

DISSERTATION

FINDING AND FIXING FRAGILITY IN MACHINE LEARNING

by

Edgar W. Jatho III

June 2023

Dissertation Supervisors:

Peter J. Denning
Armon C. Barton
Joshua A. Kroll

Approved for public release. Distribution is unlimited.

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE			<i>Form Approved OMB No. 0704-0188</i>	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instruction, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188) Washington, DC, 20503.				
1. AGENCY USE ONLY (Leave blank)		2. REPORT DATE June 2023	3. REPORT TYPE AND DATES COVERED Dissertation	
4. TITLE AND SUBTITLE FINDING AND FIXING FRAGILITY IN MACHINE LEARNING			5. FUNDING NUMBERS	
6. AUTHOR(S) Edgar W. Jatho III				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Naval Postgraduate School Monterey, CA 93943-5000			8. PERFORMING ORGANIZATION REPORT NUMBER	
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) N/A			10. SPONSORING / MONITORING AGENCY REPORT NUMBER	
11. SUPPLEMENTARY NOTES The views expressed in this thesis are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.				
12a. DISTRIBUTION / AVAILABILITY STATEMENT Approved for public release. Distribution is unlimited.			12b. DISTRIBUTION CODE A	
13. ABSTRACT (maximum 200 words) This dissertation addresses the general problem that machine learning models are fragile. Fragility arises when composing models into systems or using them in real operational environments. It also arises when model inputs are perturbed, even by small amounts, especially when perturbations are chosen by adversaries. This dissertation applies an existing state-of-the-art safety analysis methodology, System Theoretic Process Analysis (STPA), borrowed from systems safety engineering, to concrete ML applications with notable social and ethical risks to demonstrate a systematic means to argue for safe and trustworthy ML in sociotechnical systems. STPA bridges high-level goals like safety and the AI ethical principles to low level ML life-cycle design and implementation decisions. At the technical level, the dissertation introduces a novel defense for deep neural network (DNN) classifiers which exceeds state-of-the-art adversarial robustness against benchmark attacks for CIFAR-10 and CIFAR-100 datasets. The best defense, a novel stochastic, none-of-the-above (NOTA) defense, LAD-SRNA, achieves AutoAttack attack success rates less than the natural error rate in both datasets with near state-of-the-art accuracy and better than state-of-the-art robustness in classification systems. Finally, this dissertation introduces a total of 16 adaptive attacks, modifying 8 existing state-of-the-art attacks to overcome both NOTA defenses and stochastic defenses as well as a combination of the two.				
14. SUBJECT TERMS ML safety, adversarial attack defense, evasion attacks, evasion defense, adversarial examples, System Theoretic Process Analysis, STPA, system safety for machine learning, machine learning, ML, fragility, computer vision, DNN classifier, none-of-the-above defense, NOTA			15. NUMBER OF PAGES 131	
			16. PRICE CODE	
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT UU	

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release. Distribution is unlimited.

FINDING AND FIXING FRAGILITY IN MACHINE LEARNING

Edgar W. Jatho III
Commander, United States Navy
BS, United States Naval Academy, 2003
MS, Computer Science, Naval Postgraduate School, 2014

Submitted in partial fulfillment of the
requirements for the degree of

DOCTOR OF PHILOSOPHY IN COMPUTER SCIENCE

from the

**NAVAL POSTGRADUATE SCHOOL
June 2023**

Approved by: Peter J. Denning
Department of
Computer Science
Dissertation Supervisor
Dissertation Chair

Joshua A. Kroll
Department of
Computer Science
Dissertation Supervisor

Wade L. Huntley
Department of
National Security Affairs

Armon C. Barton
Department of
Computer Science
Dissertation Supervisor

Doron Drusinsky
Department of
Computer Science

Patrick McClure
Department of
Computer Science

Approved by: Gurminder Singh
Chair, Department of Computer Science

Joseph P. Hooper
Vice Provost of Academic Affairs

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

This dissertation addresses the general problem that machine learning models are fragile. Fragility arises when composing models into systems or using them in real operational environments. It also arises when model inputs are perturbed, even by small amounts, especially when perturbations are chosen by adversaries. This dissertation applies an existing state-of-the-art safety analysis methodology, System Theoretic Process Analysis (STPA), borrowed from systems safety engineering, to concrete ML applications with notable social and ethical risks to demonstrate a systematic means to argue for safe and trustworthy ML in sociotechnical systems. STPA bridges high-level goals like safety and the AI ethical principles to low level ML life-cycle design and implementation decisions. At the technical level, the dissertation introduces a novel defense for deep neural network (DNN) classifiers which exceeds state-of-the-art adversarial robustness against benchmark attacks for CIFAR-10 and CIFAR-100 datasets. The best defense, a novel stochastic, none-of-the-above (NOTA) defense, LAD-SRNA, achieves AutoAttack attack success rates less than the natural error rate in both datasets with near state-of-the-art accuracy and better than state-of-the-art robustness in classification systems. Finally, this dissertation introduces a total of 16 adaptive attacks, modifying 8 existing state-of-the-art attacks to overcome both NOTA defenses and stochastic defenses as well as a combination of the two.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

1	Introduction	1
1.1	Why Trustworthy AI Matters.	4
1.2	Tempering AI for War and Warfighters.	6
1.3	Roadmap	13
2	System Safety for ML Development and Assessment	15
2.1	Methodology	17
2.2	Case Studies	22
2.3	Discussion	36
2.4	Conclusion.	40
3	NOTA DNN Defenses and Attacks	41
3.1	Introduction	41
3.2	Background	43
3.3	Naive Adversarial Mixup	55
3.4	NOTA-Aware Attacks	59
3.5	Adaptive NOTA Defenses	64
3.6	Discussion	70
4	Improving NOTA Robustness with Bayesian Neural Networks	73
4.1	Background	74
4.2	Methodology	80
4.3	Experiments	84
4.4	Discussion	89
5	Discussion and Conclusion	91
5.1	Implications	93
5.2	Future Work	93

List of References	95
Initial Distribution List	111

List of Figures

Figure 2.1	Systems Theoretic Process Analysis (STPA) Six Step Overview. Process Adapted For ML Social and Ethical Analysis	18
Figure 2.2	Prescription Drug Monitoring Programs (PDMP) High Level Control Structure	25
Figure 2.3	Patient Care Control Loop	26
Figure 2.4	Data Decisions Control Loop	27
Figure 2.5	Notional Facial Recognition System	30
Figure 2.6	Notional CJFR Hierarchical Control Structure	34
Figure 2.7	ML Life-Cycle	39
Figure 2.8	Modeling Life-Cycle Control	39
Figure 3.1	tSNE Plots of MNIST: Light-Blue Dots on Right, Surrounding Each Class, Are AdvMix Datapoints.	56

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

Table 2.1	PDMP Risk Score System Losses	23
Table 2.2	PDMP Risk Score System Hazards	24
Table 2.3	Patient Care: Unsafe Control Actions	26
Table 2.4	UCAs: Data Decisions and Problem Framing	28
Table 2.5	PDMP Confusion Matrix (Top 1% of Scores)	29
Table 2.6	Facial Recognition System Losses	32
Table 2.7	Facial Recognition System Hazards	33
Table 2.8	CJFR Problem Framing and Data Decisions: UCAs	33
Table 2.9	Facial Recognition Model Decisions Unsafe Control Actions . . .	35
Table 2.10	Deployment, Observation and Audit Unsafe Control Actions . . .	35
Table 3.1	Model Accuracies	57
Table 3.2	Model Attack Success Rate vs. SOTA Attacks (100 Max Iterations)	58
Table 3.3	Naive AdvMix vs. NOTA Aware Updated SOTA Attacks	63
Table 3.4	Model Accuracies	66
Table 3.5	Full Table of Deterministic Models vs. NOTA-Aware Updated SOTA Attacks	69
Table 4.1	Model Accuracies	85
Table 4.2	Stochastic CIFAR-10 Models vs. None-of-the-Above Stochastic Attacks Bounds: Confidence:0, L_2 , $\epsilon = 0.5$, $L_\infty, \epsilon = 0.031$	87
Table 4.3	Stochastic CIFAR-100 Models vs. NOTA Stochastic Attacks Bounds: Confidence:0, L_2 , $\epsilon = 0.5$, $L_\infty, \epsilon = 0.031$. DET stands for Deterministic Model	88

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

AI	artificial intelligence
APGD	Auto Projected Gradient Decent
ASR	attack success rate
BNN	Bayesian Neural Network
C2	command and control
CE	cross entropy
CNN	convolutional neural network
CW	Carlini and Wagner
DLR	difference logits ratio
DNN	deep neural networks
DWBT	deterministic white box transfer
EOT	expectation over transformation
FMEA	failure modes and effects analysis
GSMA	Gradient-based Saliency-map Attack
LA	Laplace approximation
LAD	Laplace approximation with dropout
LAP	Laplace
MAP	maximum a-posteriori
MCD	Monte-Carlo Dropout
ML	machine learning
MRNA	Multi-modal Reflexive NOTA-Aware
NOTA	none-of-the-above
PDMP	prescription drug monitoring program

PGD	projected gradient descent
PRC	People’s Republic of China
SRNA	Simple Reflexive NOTA-Aware
STPA	Systems Theoretic Process Analysis
TRADES	TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization
UCA	unsafe control actions

Acknowledgments

"...as iron sharpens iron, so one man sharpens another." —Proverbs 27:17

If you surround yourself with extraordinary people, you find that you are transported to their way of thinking. If I have accomplished anything noteworthy it is only because of the extraordinary people with whom I have had the blessing of partnership. First and foremost, thank you to my amazing wife, Gwen Jatho, who is a selfless wonder, teaching our children, running our home and lives, infusing both with beautiful meaning, and enabling my unprecedented focus for these three, long/short years. To my sons, Caleb and Joshua, you are the joys of my life and I appreciate how, whenever I managed to break away for an adventure, you were ready to go! You won't know until you are a parent too how hard it was to have to say no to you sometimes when you wanted to play. To my Mom and Ken, thank you so much for your support, you were ever-ready to step in and help with the boys when needed—and you frequently did, enabling me to keep focus.

With respect to NPS, I have had the privilege of countless hours of mentorship, deep thought, exciting, meaningful work for, and alongside extraordinary thinkers. Specifically, I would like to thank, in no particular order, the following people without whom I could not have completed this task, and if somehow I had managed to, it would not have been nearly as interesting a journey or significant a product:

Professor Joshua Kroll, for countless hours, every week over three years, thank you for the tremendous investment in me, mentoring my journey to expand my intellect, sharpen my critical thinking, and broaden my perspective. I cannot thank you enough, you are a class act and continue to be an inspiration and model of academic poise. Thank you for your guidance and friendship.

Professor Patrick McClure, my own Jiminy Cricket. A strong voice for uncompromising standards, in science, work ethic, and personal conduct. Your unflinching adherence to the highest standards is an inspiration and a call to the rest of us to leave it all out on the field. Thank you for your friendship, boundless energy, constant support, and dedication to passing on your knowledge. I don't see how I could have done this without your mentorship,

collaboration and countless hours of discussion.

Professor Peter Denning, sir, thank you for the honor of working with you. I always walked away having learned about writing and Computer Science. Frequently, I also learned about life and how to live it better. Thank you for your willingness to invest in me, share your wisdom, and extend your friendship.

Professor Armon Barton, I cannot thank you enough for sharing your passion for the field of DNN defenses, and these ideas which really captured my imagination. Thank you for the many, many hours bringing me up to speed and setting me on my path. Without your mentorship, creativity, and push, I could not have done this.

To Professor Drusinsky and Dr. Huntley, thank you for your extremely valuable inputs and perspectives, your outspoken candid questioning, and ready support for the process.

To my whole committee, I believe I experienced the best combination of rigor, challenge and support I could ask for in this experience. Thank you even for the twists. I value the unique experience here, thank you for making it valuable.

Eugene Williams, thank you for all of the hours you spent as a teammate and coauthor, always faithful to the task, an extremely knowledgeable and valuable collaborator, a steady ear, and always ready with Godly wisdom. Thank you for your abiding friendship.

Logan Mailloux, you were a true Godsend! One of the fastest friends I've had, ready Godly perspective and grounded advice. I don't know how we could have accomplished the work we did without you. I don't think we could have!

Dan Lukaszewski, you paved the way for Pedro and I, thank you for your friendship, guidance, and prep through every stage.

Pedro Ortiz, we started this journey together, I always appreciated your friendship, willingness to take the extra time (that I frequently required), and dragging me back to reality when I leaped off to some unnecessarily difficult or unnecessary task.

John Fischer, Matt Litton and Chris Landis. I am glad I went first, you would all be tough acts to follow! Thank you all for your friendship, collegial spirit and for making our hall a great place to work. Time at NPS became much more enjoyable when you guys showed

up—John and Matt, I also appreciate your kindness and tolerance for how loud I am!

Finally, Rene Shelby, Andrew Smart, and Shalaleh Rismani, thank you for the opportunity to collaborate and learn from your expertise.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:

Introduction

“You can become utterly dependent on a new glamorous technology, be it cyber-space, artificial intelligence... It’ll enable you. It’ll move you forward. But does it create a potential Achilles heel? Often it does.”

—Admiral James Stavridis, USN (Ret.) [1]

Artificial intelligence (AI) and particularly machine learning (ML) systems are often unsafe. They are easily deceived and broken. They do not readily adapt to unexpected or unprecedented data. In fact, in such situations they can give confident answers when their training data and resulting algorithm have no logical basis for an answer. As a result, when ML components are placed in critical or high-consequence systems, humans often get hurt, killed or ruined. This dissertation is about how to make ML systems safer and more worthy of our trust.

Though it has recently made top headlines, this is not a new concern. In high-consequence and large-scale automated systems, the costs of accidents and unintended consequences are measured in many ways. They can be measured in the numbers of lives lost; consider Tesla autopilot deaths [2], or airline crashes [3], [4]. Some suggest they can be measured by the rise of social unrest fueled by social media algorithms [5], or the loss of individual autonomy as innocent people are detained, plead guilty, or are punished for crimes they did not commit [6], [7]. Others point to exacerbated inequalities between demographics in opportunity, services, or benefits [8], [9]. Still others point to the environmental harm that can result [10] or monetary losses that can approach the scale of a trillion dollars as in the so-called “flash crash” of 2010 [11]. If we choose to leverage ML in high-consequence domains, we must first establish that the resulting ML-enabled sociotechnical system¹ will continue to conform with society’s norms and values.

¹*Sociotechnical systems*—a phrase used in systems safety engineering, organizational development and elsewhere that recognizes that systems can and often do involve more than a collection of technical artifacts, but instead also include human social and organizational components and interactions which have equal bearing on the system’s desired performance and resilience. Whereas the word system alone can be inclusive, often instead it connotes the technical artifacts alone. We use this slightly more cumbersome phrase to purposely invoke and highlight the importance of both human and technical aspects and their interactions.

This realization has resulted in many different formulations of overarching or guiding principles for AI—from *the U.S. Department of Defense (DOD)’s AI Ethical Principles* [12] to Microsoft’s responsible AI principles [13], the European Commission’s Ethics Guidelines for Trustworthy AI [14] and the People’s Republic of China (PRC)’s “Special Action” on algorithmic governance [15]—to name but a few. However, methods that enable the repeatable validated design and implementation of systems that *operationalize* or put into effect these identified ethical principles or values remains open work.

Across the spectrum of potential and extant applications for ML, there are high-level system goals such as safety and the AI ethical principles, and there are a multitude of low-level implementation decisions which affect and effect facets of these. Yet there exists no standard or methodology to cross-walk between the two. Much of the work done in computer science circles focuses on various instrumentalizations to measure, check and correct the myriad of narrowly conceived component aspects and interpretations of these principles [16], [17]. Comparably little work has been done to investigate how such tools—and the system knowledge they reveal—can be aggregated and composed into cogent arguments that a system is safe or sufficiently reflects a principle or value.

A number of researchers have suggested that system safety engineering tools may offer a validated set of methodologies to accomplish these tasks [18], [19]. Despite these suggestions, thorough investigation into applying these tools to composing assurance arguments that an ML-enabled system is trustworthy remains largely undone. We interpret *trustworthy*, as an expert-informed assessment that a system, in its specified social and technical contexts, is an adequate solution to both its functional requirements and the problem of safety. This is with the understanding that no solution is perfectly safe, but at a minimum seeking that we can understand the risks of failure and make an informed decision with respect to the system.

Beyond the broad problem of how to compose assurance cases for trustworthy or ethical AI, there are existing serious fragility issues with ML that have no satisfactory technical solution—as would be necessary to support such an assurance case for justified confidence. One notably recalcitrant problem is adversarial robustness—an ML model’s ability to resist deception. A decade after the problem’s discovery [20], the situation is demonstrably worse, with hundreds of different attacks producing altered images, imperceptible to humans, that deceive ML classification systems. With no broad-spectrum defense or even theory for

achieving robustness, ML-enabled systems are inherently susceptible to this deception [21], [22]. The problem of adversarial evasion in DNN classifiers is a critical stumbling block to the technology's adoption in high-consequence systems. Thus, finding a way to eliminate or even reliably mitigate adversarial evasion attacks is a necessary and significant step toward any overall argument for the technology's trustworthiness in applications like self-driving cars, computer vision based identification systems or military weapons systems.

Government, industry and academia are all actively seeking repeatable, standardized, validated solutions to address these interrelated unsolved issues in ML [23]–[25]. This dissertation makes the following contributions to ML safety:

1. As a proof of concept we apply Systems Theoretic Process Analysis (STPA) in two case studies of ML-enabled sociotechnical systems and demonstrate that STPA:
 - (a) Contextualizes and identifies social and ethical hazardous system states in ML-enabled sociotechnical systems.
 - (b) Provides a logical scaffolding upon which to base social and ethical requirements to eliminate or mitigate those identified hazards either through system redesign or added social or technical controls.
2. We offer Adversarial Mixup, a new none-of-the-above (NOTA) data-augmentation-based approach to defending deep neural networks from evasion attacks. We fully evaluate the defense, advancing our understanding of NOTA defenses with the following:
 - (a) **NOTA Adaptive Attacks:** We introduce general approaches to making existing attacks *NOTA-adaptive*, then modify eight state-of-the-art attacks and restore much of their effectiveness at defeating existing NOTA defenses.
 - (b) **Adaptive NOTA Defenses:** We introduce improved NOTA augmentation strategies for training NOTA DNN models and show that the resulting models improve their performance against normal *and* NOTA-aware attacks.
3. We test a novel combination of existing (and new) stochastic approaches leveraging an underlying NOTA-defended base model.
 - (a) **Stochastic Attacks:** We adapt the same high-performing state-of-the-art attacks further to counter stochasticity by leveraging DWBT and coding EOT versions of each attack.
 - (b) **Top Performing Defense:** The combined defense results in across-the-board

state-of-the-art robustness to NOTA adaptive attacks as well as stochastic NOTA adaptive attacks in CIFAR-10 and CIFAR-100 datasets.

The above list makes a substantial contribution to trustworthy ML by demonstrating repeatable processes to build assurance that an ML system is safe and resilient to evasion attack.

1.1 Why Trustworthy AI Matters

Our society is heavily biased in favor of rapid technological advancement. We invest, look forward to, count upon, and project that technology will necessarily enhance capacity and capability. However, it is not uncommon that the opposite is true and potentially severe consequences result from a technology's premature adoption. At the other end of the spectrum, awaiting unequivocal proof of safety, is equally unacceptable, particularly when it comes to matters of national defense, international competition or war [26]. A balance needs to be struck between our drive for unbridled implementation and integration of technologies and their promised fruits—and our responsibility to preserve Western civic sensibilities—ethical treatment, due diligence, caution and safety; prioritizing human life, freedom and dignity above all. As society gains familiarity with AI, ML, and DNN tools, we are increasingly looking to them—across industry, government and even in our private lives—for opportunities to deliver solutions to our most difficult problems.

As fascinating and impressive as their achievements are, equally fascinating and yet disconcerting are the non-intuitive ways in which these technologies fail, especially in high pressure, contested or large-scale applications. Some researchers have begun to question the validity of using ML in high consequence applications [27]. While we too agree that the right answer will at times be “no,” to adopting a technology for a particular use at a particular time, we believe an open mind along with a *properly biased* [28] analysis of the technology considered in its intended context can provide a constructive set of constraints and requirements that may serve to either enable its appropriate use, or provide a detailed explanation for why not and a road-map for showing under what new circumstances it could be reconsidered.

To establish that you need a solution, you must first demonstrate that you have a problem. In the DOD our intended applications, whether in weapons systems or administrative and support systems, are certainly influential, of high-consequence and large-scale.

1.1.1 DOD Impact²

The systems Department of Defense (DOD) organizations are presently investigating or actively augmenting with ML-enabled capabilities require assurance that their subsystems are resilient to efforts to deceive. Military applications for this technology abound: analysis of acoustic and electromagnetic sensor data up and down the spectrum, machine vision, find-fix-track and targeting of adversary aircraft, subsurface, surface and land combatants, human language processing, speech recognition, voice identification, autonomous air/surface/subsurface/land vehicles, information warfare, Intelligence Surveillance and Reconnaissance (ISR) integration, robotics, cyber defense, cyber attack, tactical decision aids, strategic decision aids, and the list goes on.

Within DOD, it is well-understood that deception plays a central role in warfare. These systems must therefore be designed to be highly resilient to deception [30]. In *The Prince*, Machiavelli wrote “...although to use deception in any action is detestable, nevertheless in waging war it is praiseworthy and brings fame: He who conquers the enemy by deception is praised as much as he who conquers them by force” [31]. To Sun Tzu it is even more elemental, he is credited with having said “All war is based on deception” [32]. It is critical in national defense applications not only that systems work as advertised on the first day of combat, but also that they are resilient to the potential schemes of a crafty, equally resourced and motivated adversary.

Granted, tremendous success has been realized by ML in civilian applications and science. Although the domain of civilian industrial technological requirements has a large inherent intersection with military technological needs and requirements, it should be noted that the latter is not a perfect subset of the former. That is, the realities of warfare require that its technology must be prepared for and indeed *count upon* disinformation and deliberately deceitful actions, presentations and communications. These differing assumptions between the two domains mean that, what is primed to leverage in one, may not yet be ready in the other. Across DOD, systems that incorporate these technologies are being considered, developed and in some cases already adopted with the intent to augment or replace some of our most critical national security capabilities. In military applications, especially weapons systems and systems within the kill-chain, the problem of a lack of adversarial robustness

²This section is adapted from the coauthored NPS technical report, “Defending Against Adversarial Examples in Deep Neural Network Classifiers” [29].

must be eliminated or at least reduced and compensated for such that failures represent minimal risk. The remaining risks must be explicitly stated, known, and fully understood by warfighters and their leadership. Failing to carefully and adequately address this issue means risking the adoption of brittle technology, potentially introducing catastrophic vulnerabilities into the heart of our critical war systems.

1.2 Tempering AI for War and Warfighters³

Artificial intelligence and machine learning in particular are fragile. Though ML and other modern deep learning methods are the driving forces behind advances that make AI an important focus area today, they are distinctly vulnerable to intentional deception, manipulation and perturbation [34], [35]. One oft-prescribed antidote is human-machine teaming, i.e., the claim that by including a human ‘in the loop,’ or overseeing the task will be sufficient to counteract such vulnerabilities. However, as the history of leveraging automation has taught us, whether Three Mile Island [36], Air France flight 447 [3], or modern oversight and testing of so-called self-driving cars [37], such teaming itself is fraught and unexpectedly brittle in persistently problematic and counter-intuitive ways [38].

Such foundational fragility is not only a catastrophic flaw in warfighting systems [39], but can be a serious source of disruption and failure in administrative systems [7]. It undermines our expectations since new and seemingly capable systems appear, under straightforward evaluation, to outperform existing approaches. However, failure modes in proposed applications are often invisible [36]. Thus, AI proponents may rightly claim major technological advances, but often inadequately acknowledge or even understand the limitations of those advances. Short of a proven, repeatable methodology to identify, analyze and address hazards in such systems, we risk over-reliance on technologies that may fall significantly short of expectations, with reasonable probability of substantial harm.

Whether considering a tank target recognition system [40], an overhead sensor and intelligence synthesis suite [41] or an automated office assistant for Pentagon leadership [42], our intuition often fails to properly inform us of how best to integrate automated components, let alone validate them as safe, effective and comporting with societal and institutional values. Any

³This section is adapted from award-winning essay “AI: Too Fragile to Fight” [33] with permission from *Proceedings*.

implementation needs to proceed first with a careful examination of technological reality. Such an examination must be informed by the pitfalls and lessons learned from the past history of implementing automation in large critical-domain systems such as aviation, manufacturing, and industrial control. There are many challenges that arise in such systems and a stronger case for adoption comes with an approach that acknowledges and builds from this history.

Current AI claims are often wildly optimistic. Such claims inflate our expectations of how the resulting systems will perform, risking disillusionment when the technology fails to deliver, or worse, trust, where trust is not justified. AI cannot be considered a cure-all that applies in any or all cases, nor a product one can simply buy and implement. Rather, AI is a set of techniques that reshape problems and their potential solutions [43]. Dependable application of AI to military or national security problems must rest on concrete foundations, logical, structured arguments justifying confidence in the system.

In critical applications like military command and control (C2) or weapons find, fix, track and target systems, we must carefully evaluate new technologies against the standard of whether adopting them creates unknown, possibly more insidious problems than they solve. For large and complex problems, it is not always, or even often, the case that “any solution is better than no solution.” Rather, interventions frequently create new problems, and proponents of novel approaches have an attendant responsibility to justify confidence in them.

By highlighting a number of known deficiencies in past and current automated systems, we illustrate the difficulty in perceiving potential failures as well as motivate the necessity to evaluate the full sociotechnical system when assessing and designing a potential ML solution. In our analysis, we use a notional AI-based targeting system as a running example because target recognition is a setting where there is ongoing discussion in research, development, and policy circles. In Chapter 2, we address concerns at this higher level of abstraction. We propose adopting a proven analysis approach from system safety engineering that has been shown in other disciplines to capture the kinds of hazards and vulnerabilities we discuss below and provide a structured methodology for how to reason about and control them. The product of that process is a structured assessment that a system has sufficient controls to prevent the targeted losses. Then, in Chapters 3 and 4 we zero in on the specific issue of adversarial deception and the fragility of DNNs with respect to attack, proposing specific

techniques and defenses which can provide significant mitigation of this problem.

1.2.1 Incommensurable Goals

To begin our analysis, not only are *human recognition* and a *target recognition algorithm* not equivalent, they are not even directly comparable. They perform different tasks in different ways, and their success should be assessed using different metrics.

Human recognition in a targeting task describes not just identifying and recognizing a target, but discerning and reasoning about why a portion of a scene might, or might not, be targetable. Humans understand concepts, can generalize their observations beyond the particular situation, loosely gauge uncertainty in their assessments about target identification and can interpret novel scenarios not previously encountered with only minimal confusion. For this reason, human eyes and discernment do far more than is measured by a simple target-recognition flash card test.

“Target recognition” of an AI system is vacant by comparison. An automated vision-based classification system does far less than what is implied by the term *recognition*, a term which implicitly anthropomorphizes algorithmic systems, which simply identify and repeat known patterns, either explicitly programmed or encoded from data. Such systems cannot understand the reasons targets should be selected nor generalize beyond the specific patterns they have been programmed to handle. In a novel scenario never before encountered, it is possible that no known pattern applies. Unless that system has specific safeguards to identify and address unprecedented situations, where no data guides, AI systems will give guidance nonetheless, but without basis in knowledge of the world. The multitude of forms in which unexampled data may be encountered approaches uncountable. To be fair, humans may be perplexed as well, however, at issue is the kind of failures which occur as a result in the case of a human or an algorithm are very different, a fact which is easily leveraged by an adversary, e.g., by way of evasion attacks, as will be discussed further below.

In the real world of varying environments, degraded equipment, or where deliberate evasion and deception are expected, performance on image recognition alone does not describe performance for the extended task of target recognition.⁴ Humans are far superior at dealing

⁴Even with models trained on standard reference datasets, there is a well-documented reduction in performance when said models are tested on disjoint sets of the same provenance [44]. In contrast, human

with image *distortions* (e.g., dirt or rain on the camera lens, electrical noise in a video feed, dropped portions of an image from unreliable communications). Models trained on specific image distortions can approach or exceed human performance on a particular distortion, but the improvement does not translate to better performance on any other type of distortion [45].

Although it may be true that image recognition models can “outperform” humans on simple flash-card style tests, equating human and algorithmic performance in target selection and discernment using laboratory data or in an operational test scenario implies that performance on these tasks is comparable. This is simply false: the work being done in each case is not the same and the reliability of generated answers is vastly different. Thus, raw benchmark performance is misleading, and relying on it is dangerous.

1.2.2 Adversarial Deception: A Persistent Problem

The current best-performing ML approaches, based on deep neural networks, can seem to outperform humans on simple flash-card style qualification tests, but this performance comes at a high cost. These models do not learn general rules that apply to cases beyond the test. They extract specific patterns of the training input data and then classify according to the models trained parameters. There is no known way to determine the extent to which these patterns are training-data specific.

A particularly noteworthy illustration of this reality is the problem of *evasion* attacks, or “adversarial examples,” images designed by an adversary to confuse the technology as much as possible [34]. Some researchers have suggested that the susceptibility of ML to adversarial deception may be an unavoidable characteristic of the methods used [46]. Broadly speaking, this is not a new problem for warfighting—camouflage has existed in nature and has been practiced in an organized fashion in military units for hundreds if not thousands of years. However, to use ML effectively, we must be aware of the extent to which deception can cause misbehavior and must build the attendant doctrine and surrounding systems such that the decisions supported by the ML remain robust even when adversaries attempt to influence them.

In Chapters 3 and 4 we propose a set of methods to enable a structural change to DNN training that improves on the state-of-the-art for this important problem.

performance does not suffer this defect in performance testing between data sets [45].

1.2.3 Automation Bias

One might imagine that the problem of machine fragility can be resolved by keeping a human “in” or “on” the decision loop. That is, an ML system recommends actions to a human or is tightly supervised by a human, such that the human is actually “in control” of the outcome. Unfortunately, human-machine teams often prove fragile as well.

When guided by automation, humans can become confused about the state of the automation and the appropriate control actions to take as a result. Examples abound: in July 1988, the USS *Vincennes* accidentally shot down a civilian airliner departing its stopover at the Bandar Abbas International Airport after the ship’s Aegis system re-used a tracking identifier assigned to the plane for a fighter jet far from the ship’s position. When asked to describe the activity of the contact using the old tracking identifier, the human operator correctly indicated that it was a fighter that was descending, information that led (together with the track of the civilian airliner toward the ship) to a decision to fire on the track identified by the old identifier, leading to the disaster [47]. Although automation has improved, today human-machine team fragility has accounted for recent crashes of highly automated cars such as Teslas [2], the at-sea collision of the USS *John S. McCain* (DDG-56) in 2017 [48], and in the loss of Air France Flight 447 over the Atlantic in 2009 [3].

Several of the above instances underscore the problem of *mode confusion* between humans and machines, which can be exacerbated when information moves in complex systems or is presented with poor human factors. A related problem is *automation dependency*, where humans fail to seek out information that would contradict machine solutions. In both cases, assessing how well human-machine teams perform in context is critical, understanding whether the goal is to improve performance on average or in specific, difficult situations.

It might be argued that high overall performance or certification for operation in particular applications negates these concerns. This is also an oversimplified view, however. Let us consider the hypothetical tank target recognition system again, this time with some specific numbers. Let the system have a demonstrated 98% accuracy, and let the trained human operator have an accuracy of 88% on the same set of test scenarios. For a human operator on the battlefield, when bullets and missiles are flying and the lives of fellow Soldiers hang in the balance, will the operator question the system’s claims or will they simply pull the trigger? Can the operator trust that the machine’s better performance *in aggregate* translates

to better performance now, in their specific situation?

1.2.4 Automation Paradox

As tasks are automated away from daily practice, human operators suffer what is referred to as de-skilling [49]. So operators of our hypothetical tank system, while required to "catch" the mistakes of the system, are unfortunately less qualified to do so because they are no longer routinely required to perform the task unaided. Everyday examples abound, for instance, consider what smartphone GPS-based navigation has done for the average person's wayfinding skills: a once-routine task is now untenable for many. This phenomenon affects professionals such as pilots and even bridge watch teams.

Despite their fragility, human-machine teams can also massively outperform either humans or machines, so long as the right functions are assigned to each part and the affordances made by humans for machines and machines for humans are appropriate. Consider the game of "cyborg chess," (or "advanced chess"), in which human players use computer decision-aids in selecting their moves. In tournaments, even chess players who are weak unaided can play at a level that exceeds the world's top grandmasters and the world's top computer chess programs. Thus, human-machine integration and a focus on the processes surrounding automation can be far more impactful than the human's skill or intelligence [50].

We must not approach AI applications as self-contained artificial minds providing clear outputs fusing all features. Instead AI must be an extension of our human intelligence and organizational capability. AI is not an independent agent, but a more capable tool, applied to specific aspects of existing operations.

1.2.5 Model Security

If AI systems are fragile and subject to disruption by adversaries, must their details be kept secret? One might guess that keeping the design and parameters of an AI system secret will protect that system from being successfully attacked. Unfortunately, this does not hold true. Attacks on most types of AI systems can be performed without access to either the design or the parameters of the system. It is often possible for an adversary to guess adequate detail, extract precise model parameters by interacting with the system, or simply attempt to re-create a similar system to develop exquisite vulnerabilities or disruptive adversarial inputs

on which the system will fail [51]. Operationally, secrecy raises the bar for attackers, but not by much.

AI systems, like all military systems, must be developed, evaluated, and used with mathematician Claude Shannon’s famous maxim in mind, “the enemy knows the system” [52]. If there is an inherent vulnerability or path to system disruption, it will be discovered and leveraged. Therefore design of these systems must not only assume the adversary will be able to create an exact copy, but also that the encompassing weapons or smart sensor system must continue to function despite their access and exquisite knowledge.

1.2.6 Forging Useful Tools

Any future adoption of AI in DOD systems should include the following considerations for judicious and responsible deployment:

1. Human-machine teams must be tested and measured as a system together. Humans and machines are good at different parts of any problem set. Allocating functions and composing these capabilities isn’t straightforward, but often counter-intuitive. Careful assessment of the entire system is required to ground any claims about trustworthiness or suitability for an application. In Chapter 2, we propose an approach for how this can be accomplished in a repeatable and standardized way.
2. Demand the ruthless pursuit of what breaks the system as designed and keep an eye toward how to reformulate solutions to eliminate paths to breakage. It may be that the best solution does not involve AI or involves a less-sexy more intuitive formulation, be ready to change course and embrace it. Again in Chapter 2, we introduce a method used in other high-consequence domains to do just this.
3. AI systems must only be deployed with adequate technical and socio-technical safety nets in place. Overcoming environmental and adversarial perturbations are difficult, unsolved problems. Because AI operates based on patterns (programmed or extracted from data), its ability to operate when those patterns do not hold is inherently limited. This is an open problem we need to target for improvement, and in Chapters 3 and 4, we introduce state-of-the-art interventions to significantly mitigate this problem.

There is no denying the urgent need to move forward with AI in Navy and DOD applications. However, we need our various war-fighting communities’ eyes to be wide-open. We *require*

dependable, repeatable, validated methodologies that will analyze the full sociotechnical system and identify hidden hazards like those discussed above. AI succeeds best when it solves a clear, carefully defined problem of limited scope, supporting the existing work of warfighters or the DOD enterprise. In the following chapters we introduce tools that will help enable DOD and society at large to make these systems demonstrably safer or determine and plainly show when they are not yet safe enough.

1.3 Roadmap

Most efforts to tie safety into AI applications are focused on some combination of adhoc expert opinion, formal methods zeroed in on a particular aspect, or more narrowly on particular metrics. Absent is an overarching systemic assessment for how all of those facets can be brought together to create a unified case for assurance that a safety assessment for known and identified hazards, is trustworthy and justified. In Chapter 2 we propose adapting a proven methodology from system safety engineering to accomplish this kind of assessment in ML-enabled systems. Across two case studies, we explore whether, by applying an adapted version of STPA, we can reliably identify potential social and ethical hazards in ML-enabled sociotechnical systems that would then enable redesign of system elements to eliminate or mitigate those hazards. In Chapter 3 we focus in on a particularly recalcitrant problem in ML that threatens safety across industries and applications, particularly those susceptible to contested environments, i.e., adversarial attacks. In that chapter we investigate whether reformulating DNN classification training to include an additional *NOTA class* made up of specially-crafted data augmentation can result in broad-spectrum system adversarial robustness gains, making computer vision and other DNN classification systems safer. We also seek to better understand how such reformulation of training changes the adversarial landscape and how attacks will likely be adapted to overcome this defense paradigm. Finally, noting the gap that remained from Chapter 3 in mitigating adversarial attacks to within reasonable safety margins, in Chapter 4 we investigate whether extending deterministic NOTA models with Bayesian Neural Network (BNN) methods—like Monte Carlo Dropout, Laplace Approximation and what we believe is a first-ever combination of the two for deep neural network robustness, Laplace Approximation with Dropout (LAD), can result in synergy that provides greater broad-spectrum adversarial attack resilience. Thus both Chapter 3 and 4 seek to make contributions to a subordinate concern within overall safety in

machine learning—adversarial robustness. These techniques serve to provide a supporting assessment of safety from this particular hazardous state that would fit into the larger case for safety that an STPA analysis could provide. Finally, in Chapter 5 we discuss how these advances interrelate and mutually support the goals of making ML-enabled systems more trustworthy and less fragile.

CHAPTER 2: System Safety for ML Development and Assessment⁵

“All models are wrong. Some are useful.”

—George Box [54]

ML-driven systems promise new automation capabilities and provide valuable new tools across a diversity of applications. At the same time, questions abound as to these systems’ trustworthiness and safety. For this reason, stakeholder communities across academia, industry and civil government have all proffered interventions designed to build the ability to make reliances on ML-driven systems for particular purposes. However, to-date, the bulk of these interventions are *instrumental* changes at the *component level*, which aim to improve particular performance metrics e.g., “fairness metrics.” One often-cited taxonomy of AI safety issues [17] has been critiqued as taking an overly-instrumental frame [19], a critique we extend to follow-on work focused on ML [55]. There is a substantial and long-recognized gap between improved component performance and systemic safety [36]; instrumental, component-wise improvement will not enable the safe use of ML.

Other efforts at intervention, i.e., management, policy and design frameworks or lists of ethical principles, while taking a wider view of the problem, are often a challenge to extract implementable requirements from or bridge to actual technical decision-making. Closing that divide, between high-level policy goals such as ethical principles, and practical low-level technical decisions, is a key enabling function required to address legitimate concerns about ML solutions. This is particularly true as new regulatory demands drive a need for assurance against failures in high-risk applications [56], [57]. Metric-based evaluation on its own, represents a narrow view of model performance, particularly when considering social and ethical risks. It fails to address the more general critical equities at stake [58], [59]. Fairness metrics, a common proxy for identifying social and ethical concerns, nevertheless fall short of adequately modeling underlying human values [60]. Finally, even when a social or ethical hazard related to a ML system is widely recognized, it can be difficult to assign responsibility,

⁵This chapter was developed from a research project on which I was lead author [53]; coauthors included Professor Logan Mailloux, Mr. Eugene Williams, Professor Patrick McClure, and Professor Joshua Kroll.

or to determine appropriate interventions to mitigate or prevent the hazard [61]–[64].

In response to these shortfalls, several ambitious efforts aim to create practical, effective evaluation frameworks for identifying harms and to mitigate social and ethical risks. For example, the U.S. NIST’s AI Risk Management Framework [57] and the pending “AI Act” legislation in the European Union both categorize concerns about harm resulting from ML-driven systems as a risk management problem and envision solutions in standardized evaluation frameworks. However, these frameworks rely on consensus best-effort and expert judgement; assessments of social and ethical risks remain ad-hoc even if systematized. Instead, well founded risk governance should proceed from formal assessment, rigorous analysis, and procedural validation, realities recognized by ML practitioners and academics alike [25], [65]–[67].

Systems safety engineering provides well-studied tools, techniques, and procedures to identify and control risks in complex sociotechnical systems (collections of interacting organizational, human, and technical components) [68], [69]. Although others have suggested that safety frameworks can perform similarly for ML system risks, including social and ethical risks [18], [62], [70], the concrete use of these techniques, along with any investigation of whether they can surface social and ethical risks in particular, remains unexplored. To test the hypothesis, we apply Leveson’s System Theoretic Process Analysis (STPA) [71] to two representative ML systems, giving special attention to socially and ethically problematic outcomes, operationalized as safety hazards.

Safety analysis is a tool for identifying risks so their presence can be acknowledged and accepted or mitigated through additional controls or redesign that eliminates their possibility. Although in some cases, it is desirable to use the results of safety analysis to create *assurance cases*, arguments to a skeptical expert that a system will not manifest undesired behaviors, safety analysis cannot produce formal guarantees or even claims of completeness with regard to the identification of risk.⁶ While we cannot claim to identify all risks, our goal is to demonstrate that these methods support the coherent composition of component-level interventions and that they outperform unstructured analysis either by discovering more risks or discovering risks with less effort.

⁶Indeed, even formal methods such as verification cannot produce such guarantees, relying on assumed bounds or on epistemic assumptions about the relationship between invariants and reality.

An advantage offered by system safety engineering frameworks as a whole, is their ability to bridge abstract safety policies to implementable requirements, which are often difficult to achieve or operationalize solely through technical means [65]. Tools from this discipline further have the advantage of being regularly applied in consequential domains, being well studied, and providing a strong basis on which to systematize efforts to identify social and ethical risks [72], [73]. Component-level methods include traditional safety-through-reliability techniques like fault-tree analysis (FTA) [74] and Failure Mode and Effects Analysis (FMEA), both quantification-oriented approaches used for decades to reduce the number and frequency of failures in systems under analysis [75]. Leveson's Systems-Theoretic Accident Model and Process (STAMP) body of tools [69], however, explicitly rejects the notion that reducing failures necessarily improves safety, noting that safety is a property of systems, *not components*. STAMP models sociotechnical systems in terms of control and control structures. In this paradigm, unwanted losses and hazardous system states are understood as the result of insufficient control within the system.

Chapter Contributions:

- Demonstrate the utility of applying STPA to identify social and ethical risks in ML-enabled systems.
- Offer a standard, repeatable bridge between low-level ML design decision-making and high-level social and ethical values and consequences.
- Advance the argument that ML trustworthiness must be established as a system property.
- Provide a reference model for the community to study and expand.

2.1 Methodology

Historically, various methods for identifying and mitigating risks in complex systems sought to do so using two general approaches: 1) reducing the probability of component failures, and/or 2) reducing the severity of the effects of each component's failure. The logic being that, first, lower failure rates of constituent components should necessarily lead to lower system failure rates and, second, lower severity of component failure, often through the introduction of redundancy, should likewise lead to less severe overall system failures. Some systems, however, are so complex or tightly coupled that even in the presence of (and not

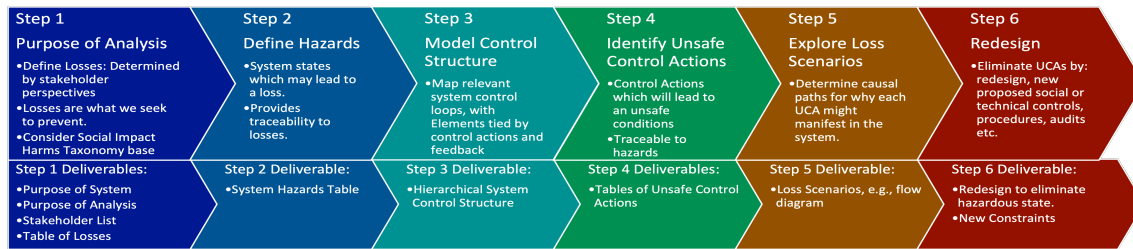


Figure 2.1. STPA Six Step Overview. Process Adapted For ML Social and Ethical Analysis. Adapted from [71].

uncommonly *because of*) redundancies and fail-safes, they can and do experience unforeseen catastrophic interactions between simultaneous component failures that, though estimated to be of vanishingly small probability, in fact commonly co-occur. Far from being unlikely, accidents of this kind have been deemed inevitable, or even *normal accidents* [36]. In his seminal work of the same name, Perrow points out that component failures or human errors are *not required* for a failure or accident to occur, but instead system failure may result from unforeseen interactions between *properly functioning components* [36]. Such interactions and subsequent failures are *emergent* system behaviors, arising from inadequate or unsafe control actions in a system's control structure [69], [76]. Examples of this phenomenon abound: consider the Ariane 501 explosion [77] or the Mars Polar Lander crash [78], both driven not by component failures but by control failures far from the proximate causes of loss. Methods that only consider components and their failures *cannot* foresee system failures of this kind.

2.1.1 Six-Step STPA for ML Social and Ethical Impact

We compose STPA into six steps as shown in figure 2.1 for clarity. While STPA is normally defined by four steps, this minor restructure enables our analysis to leverage deliverables more naturally to inform discussion and decisions when considering how to redesign or build-in new social and ethical constraints. By considering the full sociotechnical system, our STPAs contextualize low-level development decisions and their resulting ML hazards with respect to identified social and ethical losses. Recognizing and responding to the social and ethical risks of an ML model requires viewing the system in its context of use, as part of its broader sociotechnical system and in light of its control structure [79].

Define the Purpose of Analysis

Walking through the STPA method, step one begins by defining the purpose of our analysis: defining stakeholders, understanding the intended system and its existing sociotechnical context, setting the scope of analysis, and understanding the kind of losses we want to prevent. Under STPA, *losses* involve the deprivation, ruin or forfeiture of something valued by stakeholders; in this case we cast losses in terms of societal or individual harms. At this stage it is important to understand that losses can, and frequently do, exist outside the scope of the system of interest (SOI). However, STPA is designed to address losses wherever they occur, and modeling when losses occur beyond the system boundary can help define new system requirements to control the effects of the loss or may suggest broadening the scope of analysis in order to effectively control the loss.

Define System Hazards

In step two, we capture and define the system's hazardous states. *Hazards* are system states which could lead to a defined loss. In each analysis, we created a substantial initial hazard list, and revised it during each later step of STPA by consolidating and refining hazards. It is relatively easy to generate a large list of potentially hazardous systems states, so the analysis team must constantly be reminded to only go as low in abstraction as necessary to reason about controlling hazards. High-level hazards may be decomposed into more detailed hazards as the model is enriched to lower levels of abstraction. The output of this step is a complete list of system hazards with respect to the defined system losses.

Model Control Structure

Next, in step three, we use system documentation, requirements, organizational structure and other contextual information to build out high-level mappings of the system's control structure with linkages to important contextual elements, see figures 2.6 and 2.2 for detailed examples. In this step, control loops are modeled where the controller and its controlled process are tied together by *control actions* which are executed by a controller and *feedback* provided by measurements of the controlled process. For the purposes of our case studies, we scoped each analysis to focus on two or three of the most interesting ML-related control loops. The output of this step is a full system control structure, situated in necessary context and specified to the necessary level of abstraction to enable system control analysis.

Identify Unsafe Control Actions

In step four, our STPA analysis proceeds by discovering *unsafe control actions* (UCAs) in each control loop. These UCAs are found by identifying necessary control actions between the controller and the controlled process and then applying a slightly modified (as described in Section 2.3) STPA rubric to uncover paths to hazardous system states. That is, for each control action we ask whether any of the following cause the system to enter a hazardous state: not providing the control action; providing the control action; executing the control action too early, too late or in the wrong order; or providing the control action to excess or insufficient degree. Using this rubric, we build the output for this STPA step, a UCA analysis table which identifies all of the potential hazardous system states which may result from a misapplied control action. See 2.8 for an example of a completed UCA analysis table.

Explore Loss Scenarios

Next, in step five, we examine loss scenarios, which are causal paths leading the system into hazardous states which may result in a loss. A number of loss scenarios are proposed and then walked through the control structure to examine why an unsafe control action might occur or why an otherwise safe control action might be improperly executed or not executed, verifying that a UCA can potentially lead to a hazard. Detailed analysis at this step ensures measurable criteria are attached to each control action and feedback loop. Additionally, this step serves as a verification check of the UCAs to ensure the analysis is complete and each UCA is correctly mitigated.

Redesign

Finally, in step six of the STPA we are able to leverage the outputs of the previous steps to trace stakeholder defined social and ethical losses back to various competing or contributing control actions, (read as low-level technical decisions), providing a full social and ethical context within which to consider and balance these controls.

Once complete, the outputs from each step in STPA enable traceability from every stakeholder defined loss to each hazardous system state to each verified contributing UCA. The STPA process not only identifies system hazards, but its outputs link and contextualize them in a way that leads directly to socially and ethically informed development and operational decision-making. Moreover, these new requirements or constraints are solution neutral: they

express the need for better control but not explicitly how to effect that control. UCAs can be addressed in the most effective and efficient way suited to the particular system. For example, in some cases it may be possible to redesign the system to eliminate a risk entirely, while in other cases it may be desirable to instead re-balance existing control options and design decisions and institute feedback to monitor future risk. Other situations may require the provision of supplemental training to operators to avoid risks. Another possibility is that the STPA reveals that insufficient control is possible with available technology, i.e., the incumbent risks with a particular ML component cannot be mitigated to an acceptable level and it is necessary to instead forego fielding the system. Even in such a scenario, however, the outputs of the analysis—the loss list, the hazards, safety control structure etc.—provide a structured contextual knowledge-base with which to reason about and identify necessary requirements to reconsider that decision.

2.1.2 Team Composition and Study Limitations

Our first case study team was composed of five ML researchers, one sociologist, and one STPA subject matter expert. For the second case study, our team was composed of four ML researchers and one STPA expert. Four members of the team participated in both analyses, including the primary researcher and our STPA expert. In addition to working with an experienced STPA expert, each member of the team worked to become thoroughly familiar with STPA by reading core STPA expositions – Leveson’s *Engineering a Safer World* textbook [69], studying the *STPA Handbook* [71], and watching training videos [80].

In the first case study we did not have direct access to the prescription drug monitoring program (PDMP) scoring system or its datasets. However, we were able to derive necessary details from publicly available operating manuals, pharmacy work-flows, state legal references, and training documents. Additionally, we rely on recent research, which re-creates and tests ML models using similar datasets and features as are known to be used in real-world systems. State manuals describing key variables as “predictive of unintentional overdose death” provided insight to understanding model optimization targets and extracted features [81]–[84].

Likewise, we did not have direct access to a law enforcement version of the subject facial recognition system or its datasets. However, we were able to derive necessary details from patents, company descriptions, legal requirements, and government testimonies. Additionally,

we were able to study open source API documentation and were granted access to a publicly-available version of the system to gain further insight. Given the top-down framing of STPA, these artifacts proved sufficient to provide ample sociotechnical context, operational details, and developmental insights to perform an effective STPA for social and ethical impacts on each system.

Due to the nature of safety analysis, there are no *guarantees* of safety or analytical completeness. Approaches that have offered such guarantees fail to account for the reality of unknown unknowns [28]. What we can arrive at, are necessary elements of safety and reasonable confidences in the assurances to those specific ends. The intent of this study is to be a pilot to demonstrate proof of concept, not act as a full-scale hazard analysis. We leave it to future work to perform comparison studies with previous methods. Although our analysis can demonstrate problematic dynamics between the ML components and the rest of the system, we did not have a way to test or experiment with our proposed interventions. Therefore, any solution we recommend is necessarily hypothetical. A fully integrated team in a design effort using these methods would not face most of these limitations.

2.2 Case Studies

2.2.1 PDMP Risk Score

Prescription Drug Monitoring Programs (PDMPs) are mandated in all 50 states and are intended to prevent or curtail widespread healthcare issues such as drug addiction, misuse, and overdose deaths [85]. A majority of these programs require physicians, pharmacists or their staffs to review a patient's risk scores prior to writing or filling prescriptions for certain schedules of drugs [82], [86]. Risk scores are calculated by ML systems trained on a variety of data sources [83], [87]. Thus, these ML-enabled PDMP systems assist in governing the health care of hundreds of millions of people across the United States.⁷

⁷The author performed an informal search of state public health websites and official news releases and confirmed 26 of 51 states and the District of Columbia use an ML-based scoring system as a major component of their PDMP. Additionally, five of the top seven U.S. pharmacy businesses (by 2021 prescription revenue) [88] require it.

2.2.2 PDMP ML STPA Analysis

We begin step one of our PDMP risk score STPA by defining the purpose of our analysis. The purpose of this STPA is to identify and remediate potential downstream sociotechnical harms that result from, or be contributed to by, the PDMP risk score. We identified major stakeholders for the system as patients, doctors and pharmacists. Leveraging a recently published algorithmic harms taxonomy [89] and the process described in the STPA Handbook [71], we initially identify 30 distinct potential downstream social and ethical losses (See appendix ?? for complete list), which we pare down to five general loss categories (listed here in table 2.1).

Table 2.1. PDMP Risk Score System Losses

PDMP Losses	
1	Death, Injury or Disability (i.e., Patient Death, Untreated Pain, Additional Physical or Mental Injury)
2	Disparity of Benefit/Harm (i.e., Allocative, Representational or Quality of Service Disparities)
3	Social/Economic Injury (i.e., Damage to Reputation, Occupational Damage, Family Damage, Privacy Violations)
4	Damage to Quality of Healthcare (i.e., Abandonment, loss of autonomy in clinical judgement, reduced accessibility to doctors)
5	Coerced Criminality or Unsafe Treatment (i.e., Herding to Unsafe/Illicit Behavior, Increase in Law Enforcement Scrutiny)

In step two, we identify 11 hazardous system states and map each to the losses they potentially contribute to, shown in table 2.2. While losses are not confined directly within the system of interest (SOI) and may exist far downstream, hazards *are* confined within the SOI but can still manifest at various levels within its hierarchical control structure. Observe that some hazards are specific to the ML model, whereas others occur at higher levels of control and may even be contributed to by low-level hazards. For example, over-prescribing, a system hazard manifesting between patient care and patient health—describing where opioids are prescribed to someone who in fact is likely to become addicted or overdose—may result from the low-level hazard where the PDMP risk-score exhibits excessive false negatives. Conversely, under-prescribing is a system hazard which arises at the same level of abstraction as over-prescribing, and can result from the opposite low-level hazard, excessive false positives.

Table 2.2. PDMP Risk Score System Hazards

PDMP Risk Score Hazards						
	Hazards	L1	L2	L3	L4	L5
1	Over-prescribe	X			X	
2	Under-prescribe	X		X	X	
3A	Inappropriately Scored - High	X		X	X	X
3B	Inappropriately Scored - Low	X			X	
4	Score Leaked	X	X	X	X	X
5	Problematic/Biased Data		X		X	
6	Abandonment	X		X	X	X
7	Not provided most effective treatment.	X		X	X	X
8	Patient gives up on medical system.	X		X		X
9	Excessive false positives.	X	X	X	X	
10	Excessive false negatives.	X			X	

Next, in step three, we derive the hierarchical system control structure (Figure 2.2). This effort involves a great deal of analysis to determine proper scoping, system boundaries and necessary levels of abstraction. Indeed, we found it most productive to model the system with two control structures of interest. First, an operational control structure, where the system is being employed and which yields a rich understanding of the sociotechnical context and resulting harms. The second, is a development and implementation life-cycle control structure, where critical low-level development decisions are made. Modeling and adapting the ML-lifecycle control structure yields a way to model low-level ML decisions as *control actions*, with the selected metrics of social and ethical concern serving as feedback⁸. See figure 2.4 for the Problem Framing and Data Decisions control loop. For the remainder of this case study, we limit our analysis to the two most impactful of the seven primary control loops in the PDMP risk score control structure, namely the operational loop *Patient Care* (Figure 2.3) and the development loop, *Data Decisions and Problem Framing* (Figure 2.4).

In step four we identify unsafe control actions (UCA). These are captured using a rubric from STPA. For each identified control action, we consider whether providing, not providing, or mis-applying the control action (either too early, too late, in the wrong order, too long or too short) can result in hazardous states. Each activity that results in a UCA is recorded in the table linked to the hazardous system states which can result.

In step five, we consider loss scenarios, situations where unsafe control actions can arise.

⁸As will become more apparent in the CJFR case study, even the decisions regarding the social and ethical measurements and monitoring are themselves control actions derived in the *Data Decisions and Problem Framing* control loop.

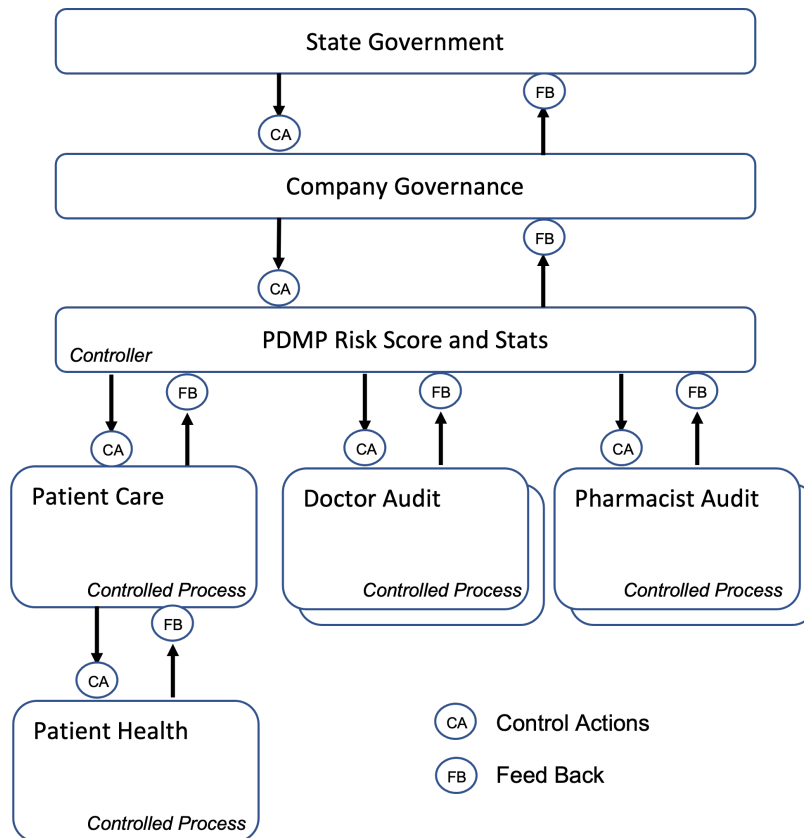


Figure 2.2. Prescription Drug Monitoring Programs (PDMP) High Level Control Structure

This step provides insight into how UCAs may be eliminated, mitigated or constrained. Identifying loss scenarios involves "walking" the UCAs and discovering how each may arise. This process reveals how some UCAs may even coincide, with overlapping resulting hazards and losses. The process then structures the tensions between control actions with respect to social and ethical losses, i.e., with respect to ML, capturing and contextualizing tensions between low-level ML decisions which contribute to stakeholder losses.

To demonstrate, we choose the control action “define the problem,” considering how this control action may fail to execute, or be improperly executed. One conceivable avenue to this end might be the choice of an inappropriate proxy to the task—a mismatch between a chosen proxy and the system goal [8], [90]. This can be caused by stakeholder beliefs that a proxy is a good representation of the problem, beliefs which follow from inadequate analysis

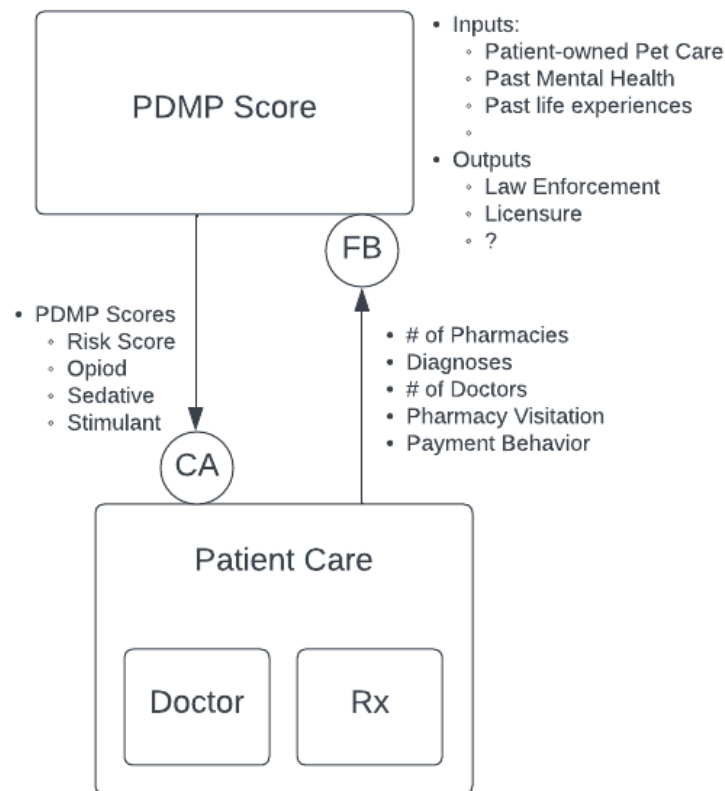


Figure 2.3. Patient Care Control Loop

of the proxy’s validity. During the course of the PDMP risk score STPA it became clear that, though the ostensible aim of the risk score is to reduce overdose deaths and opioid use disorder (OUD), the data features leveraged in its determination coincide instead with identifying drug diversion. Thus *drug diversion* is the chosen proxy for the PDMP risk score. Our question—is this modeling valid?

Table 2.3. Patient Care: Unsafe Control Actions

UCAs: Patient Care Control Loop					
Control Action	Not Provided	Provided	TE TL	Too Low	Too High
Risk Score	Score defaults to zero [H1]	[H6,7,8]	N/A	[H1], [H3,10]	[H2,3A, 6,7,8,9]

Finally, in step six we look for opportunities to apply constraints in the form of redesign or

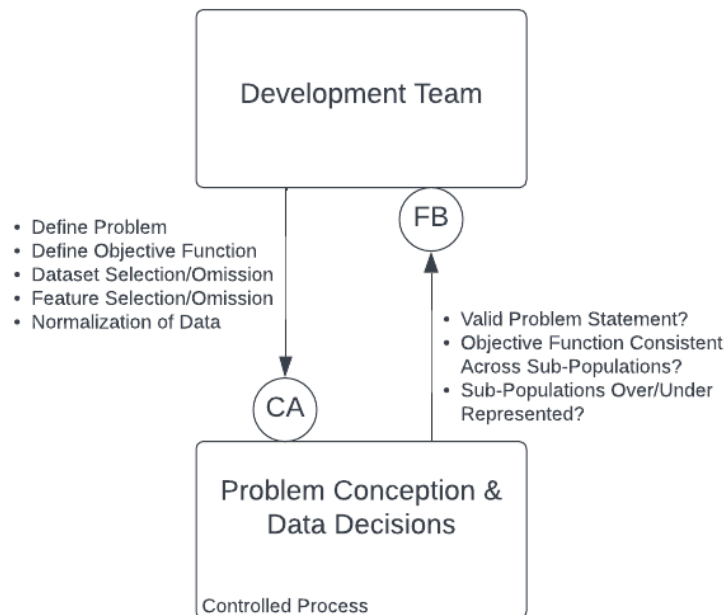


Figure 2.4. Data Decisions Control Loop

adding new social or technical controls to the system to eliminate or mitigate the hazardous states. For example, in the loss scenario above regarding whether a proxy is valid, one could propose validity checks for potential proxies as a constraint. Several resources are available for implementing such a control. Jacobs and Wallach suggest testing for various forms of construct validity or reliability such as face validity, content validity and convergent validity [90]. Whereas, Coston et al suggest a proxy can be evaluated along three lines [27]: 1) *validity*, whether it closely reflects or assesses the concept that the designer intends, 2) *value alignment*, whether proxy goals and behavior can be made to comply with collective stakeholder values, and 3) *reliability*, the extent to which the proxy's output is repeatable, consistent and stable.

Table 2.4. UCAs: Data Decisions and Problem Framing

UCAs: Data Decisions and Problem Framing Control Loop					
Control Action	Not Prov	Provided	Too Early/Late	Too Few/Little	Too Many/Much
Define Problem	N/A	H5, H3	N/A	N/A	N/A
Define Obj. Function	N/A	H1-3, H5-10	N/A	N/A	N/A
Dataset Selection or Omission	N/A	H1-3, H5-10	N/A	N/A	N/A
Feature Selection or Omission	N/A	H1-3, H5-10	N/A	H1-3,H6-10	H1-3,H6-10
Data Normalization	N/A	H3, H5, H9, H10	N/A	H3, H5, H9, H10	H3, H5, H9, H10

PDMP Findings

In systems like PDMP, which affect lives, health and livelihood at scale, it is incumbent on developers, company management and government officials to demonstrate due diligence before fielding a system [58], [91], [92]. We contend this can be done by providing STPA-generated evidence that an ML-enabled system not only improves the performance of the system which it is replacing, but that it does not also introduce new unacceptable negative social and ethical impacts.

As discussed above, one important fact that our analysis highlighted was the gap between stated goals for the system (reducing opioid use disorder and preventing overdoses) and the optimization target for the score (minimizing drug *diversion*, the illegal distribution or abuse of prescription drugs or their use for purposes not intended by the prescriber). The use of this proxy measure depends on the assumption that reducing drug diversion saves lives; however, we conclude that the opposite relationship could be an unintended outcome of the system by considering the hazards, control loops, and UCAs we identified in analysis. Indeed, nationwide statistics show: 1) a marked reduction in drug diversion and opioid prescribing coinciding with the adoption of PDMPs in every state; and 2) an increasing overdose death rate overall [93]. Although we discovered a proxy mismatch was previously noted for this system, STPA analysis made the issue apparent early in the process using only design-level instantiation, i.e., defining the purpose of the analysis and considering the potential social harms. By contrast, prior identification of the problem required substantial research efforts and collection of data about fielded systems [94].

Second, the STPA control loops reveal mechanisms whereby actions taken by prescribers and dispensers to reduce their business risk (e.g., lowering their prescribing average milligram morphine equivalents to within norms for their specialty), have the potential to drive patients to more desperate and dangerous behavior as they seek the treatment no longer offered

by their physicians. These observations are validated in literature on PDMP [95], [96]. Provider and dispenser "report cards" [97], [98] and unsolicited PDMP reports sent to state licensing boards, and law enforcement agencies [99] can be seen as control actions that drive provider behavior—provider per patient *milligrams morphine equivalent* compared with their specialty’s average along with other PDMP statistics create pressure on physicians and pharmacists which affect individual-level care decisions [100], [101]. Whether this results in discontinuing treatment for pain or termination of care, both have been documented [94].

Finally, our STPA highlights the social and ethical losses introduced by the PDMP risk score that can result from high false positive and high false negative rates. These are indeed concerning given research which finds that systems based on the same datasets and features used in the most ubiquitous PDMP risk score systems, were shown to have significant false positive and false negative rates (see the appendix for the confusion matrix in table 2.5) [102]. High error rates exist even when considering the patient population with the highest 1% of

Table 2.5. PDMP Confusion Matrix (Top 1% of Scores). Adapted from [102].

		Prediction outcome		total
		p	n	
actual value	p'	TP= 7,640	FN= 13,231	P'
	n'	FP= 61,255	TN= 6.8+MILL	N'
total		P	N	

risk scores. Of this group, only 11% are true positives. In the same top-1%-by-risk-score population, the model would only identify 37% of the total true positive patients who need intervention [81], [87]. For any ML system whose effects are felt at scale, two take-aways follow from this analysis. One, it is relevant and important to consider how often identified hazardous states are entered. Two, designers should structure the surrounding sociotechnical system to compensate appropriately in order to mitigate these foreseeable losses. STPA provides a systematic means for identifying and capturing these seemingly obscure facts and coupling them with measurable and auditable mitigation and elimination strategies. In this

way, STPA can serve as a means for providing a standard for the minimum necessary due diligence to field ML-driven high-consequence sociotechnical systems.

2.2.3 Criminal Justice Facial Recognition System (CJFRS)

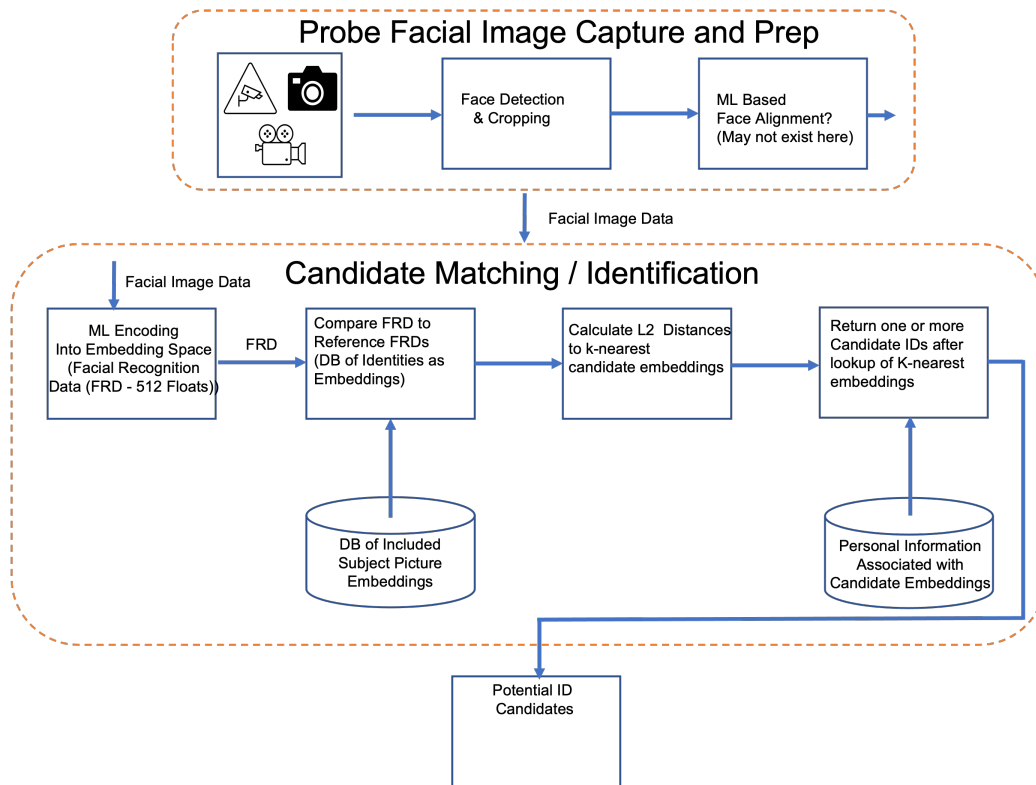


Figure 2.5. Notional Facial Recognition System

As of 2016, over 3,200 state and local law enforcement agencies reported leveraging facial recognition technology to assist in criminal investigations, and by 2020 at least 20 federal agencies reported using face recognition for law enforcement purposes [103]. In testimony before the U.S. House Oversight and Reform Committee, it was reported that the FBI Criminal Justice Information Services division received 152,565 facial recognition search requests to its NGI-IPS repository⁹ between December 2017 and April 2019 [104]. This total averages to almost 9,000 facial recognition identification requests per month. However, at this time there is no standard U.S. guidance for the integration and use of AI or ML-based

⁹Next Generation Identification - Interstate Photo System.

facial recognition technology in law enforcement settings [105].¹⁰

Currently, there is no overarching governance process or standardized guidance for the use of ML-driven facial recognition algorithms in complex sociotechnical law enforcement structures. This is troubling given the number of agencies using these types of technologies as inputs to life-impacting, consequential decisions [106]. As a result facial recognition policies, procedures and implementations vary within law enforcement agencies. In the best case, a trained facial recognition team regularly probes the database and verifies results; in other deployments, lone untrained local law enforcement investigators use a facial recognition systems with little or no training [6].

Lastly, one industry-leading U.S. company provides the contracted support to the FBI, U.S. Immigration and Customs Enforcement, and the Department of Defense. Moreover, they boast a facial recognition database of over 30 billion photos—representing the equivalent of nearly four photos for every human on earth [107], [108]. These photos are harvested from the internet, social media, news articles, company websites, and other available data sources. As of 2016, it was estimated that half of all American adults were already in law enforcement facial recognition systems. In querying these photos, the above company also reports an astounding “99%+ accuracy across all demographics” when seeking to match photos to identities according to NIST testing standards [109].

CJFRS STPA Analysis

We begin our STPA of the ML-enabled criminal justice facial recognition system by defining the system purpose and the purpose of our analysis. We take the purpose of the CJFRS to be to take an input probe photo of an unknown person suspected of a crime and return k number potential matches. This system is intended to be used as an aid in investigations in criminal justice. The purpose of our STPA is to identify and when possible address potential downstream sociotechnical harms that might result from or be contributed to by the CJFRS. Major stakeholders are individual citizens, society as a whole, law enforcement officers, prosecutors, defense attorneys and the algorithm’s owning company. Because of the complexity involved with this ML-driven sociotechnical system, we found it helpful to categorize losses into two broad types: individually-experienced losses and losses that

¹⁰The FBI reports having its own standard procedures for using facial recognition in investigations. Training includes manual reviews of matching probe and candidate photos. [104]

Table 2.6. Facial Recognition System Losses

Facial Recognition System Losses	
Societal Losses	
S1	Loss of Justice • Wrongful Convictions or No Convictions • Loss of Confidence in Overall System by Population or Demographic • Loss of Public Service Reputation
S2	Loss of Public Service Resources (i.e. Manpower, Money, Time, Equipment)
S3	Loss of Opportunity to Population or Demographic
S4	Loss of Privacy: Surveillance State or Perception Thereof
Individual Losses	
I1	Loss of Life / Physical Harm / Harassment
I2	Loss of Opportunity (e.g. Missing out on College, Scholarship, or Job due to Wrongful Arrest)
I3	Loss of Resources (Time/Money/Effort)
I4	Loss of Confidence in Justice System
I5	Loss of Privacy / Autonomy - Surveillance

manifest at the societal level, as seen in table 2.6. We found this categorization useful to carry through the STPA process, separating hazard analysis and the consideration of loss scenarios as they dealt with different parts of the larger CJFR system.

Next, in step two we identify 21 hazardous system states across all levels of abstraction from the prosecutor’s office down to the actual facial recognition algorithm and database. We then map each hazard to one or more of the related losses (Table 2.7, below). Again, a lower level hazard may propagate up the layers of abstraction to a higher level system hazard.

Proceeding to step three, we model the system control structure, deciding on a bilateral design separating the operational control from developmental control, each side “controlling” the facial recognition algorithm at bottom, figure 2.6. On the development side, STPA’s structure enables reasoning about machine learning life-cycle decisions. As now explicitly shown in the control structure, we model the development team as a controller and the three

Table 2.7. Facial Recognition System Hazards

Facial Recognition System Hazards										
	Hazards	S1	S2	S3	S4	I1	I2	I3	I4	I5
1	Wrongful Warrant Issuance	X	X			X	X	X	X	X
2	Investigator Cognitive Bias (i.e. automation bias, anchoring)	X	X	X	X	X	X	X	X	X
3	Leaked Data (Identities)		X			X	X	X	X	X
5	No Guidance on FR use. Lax Control of FR confidence level - Low	X	X		X	X	X	X	X	X
6	Investigator Improper Confidence in Results	X	X	X		X	X	X	X	X
7	Face Detection Evasion	X				X	X	X	X	X
8	Unverified Probe Photo Alteration (i.e. Photoshop)	X	X			X	X	X	X	X
9	Operator Error / Training Deficiency	X	X	X	X	X	X	X	X	X
10	Unauthorized/Unintended/Unethical use of FRT (i.e. surveillance)	X	X	X	X	X	X	X	X	X
11	Actual Perpetrator Not in Database	X	X	X		X			X	
12	False Match (False Positive)	X	X			X	X	X	X	X
13	False Non-Match (False Negative)	X	X			X			X	
14	Different Performance of FR Between Demographics	X		X	X					
15	DB Tampering (Ingested Photos)	X	X	X	X	X	X	X	X	X
16	Frame an Individual through Matching Evasion Attack					X	X	X	X	X
17	Sparse Photo Representation for Identity	X	X	X		X				X
18	Mislabeled Photos in DB	X	X			X	X	X	X	X
19	Identity Mismatch in DB					X	X	X	X	X
20	Face Not Detected or Wrong Portion of Image Captured as Face	X			X					X
21	Reinforcing Classification Feedback Loops			X						
22	Stale DB Images or Embeddings of a Subject(s)	X	X			X	X	X	X	X

life-cycle phases as controlled processes: *Problem Framing and Data Decisions* (PFDD); *Modeling Decisions*; and *Deployment, Observation, and Audit*. In this framing, *decisions* made by the development team represent control actions.

Table 2.8. CJFR Problem Framing and Data Decisions: UCAs

UCA 1 - Problem Framing and Data Decisions				
Control Action	Not Providing	Providing	Too Early/Too Late (Wrong Timing)	Too Low (TL) / Too High (TH)
(1) Define Problem	Ill-defined problem [H6,10,14,17,20]		Defining the problem after a model is built [H6,10,14]	
(2) Define Social and Ethical Success Metrics	Model behavior and desired behavior mismatch [H1,6,7,10,12,13,14, 20]	Insufficient success metrics [H1,6,7,10,12,13,14, 20]		
(4) Database Update Policy	Stale model [H6,11,12,13,17,22]	Updating with bad data [H7,11,12,13,14,15,16,21]		Inappropriate data update rate can lead to undesirable model behavior [TL: H6,12,13,22]
(5) Source Weighting	Trusted data supplanted with untrusted data [H15,16]			
(6) Data Processing	Inappropriate data pre-processing and balancing [H11,12,13,14,17]	Pre-processing may introduce additional errors [H12,13,14]		

In step four we focus on the development control structure vice the operational side, finding a combined 32 unsafe control actions (UCAs) as shown in tables 2.8, 2.9, and 2.10. Our goal in focusing on these concerns is to structure reasoning about problems that arise from decisions there and to monitor and adjust development decisions to mitigate (or in the best case, eliminate) social and ethical system hazards. Again, the benefit of framing the process in this way is that it casts every decision in the development cycle in terms of its effects on

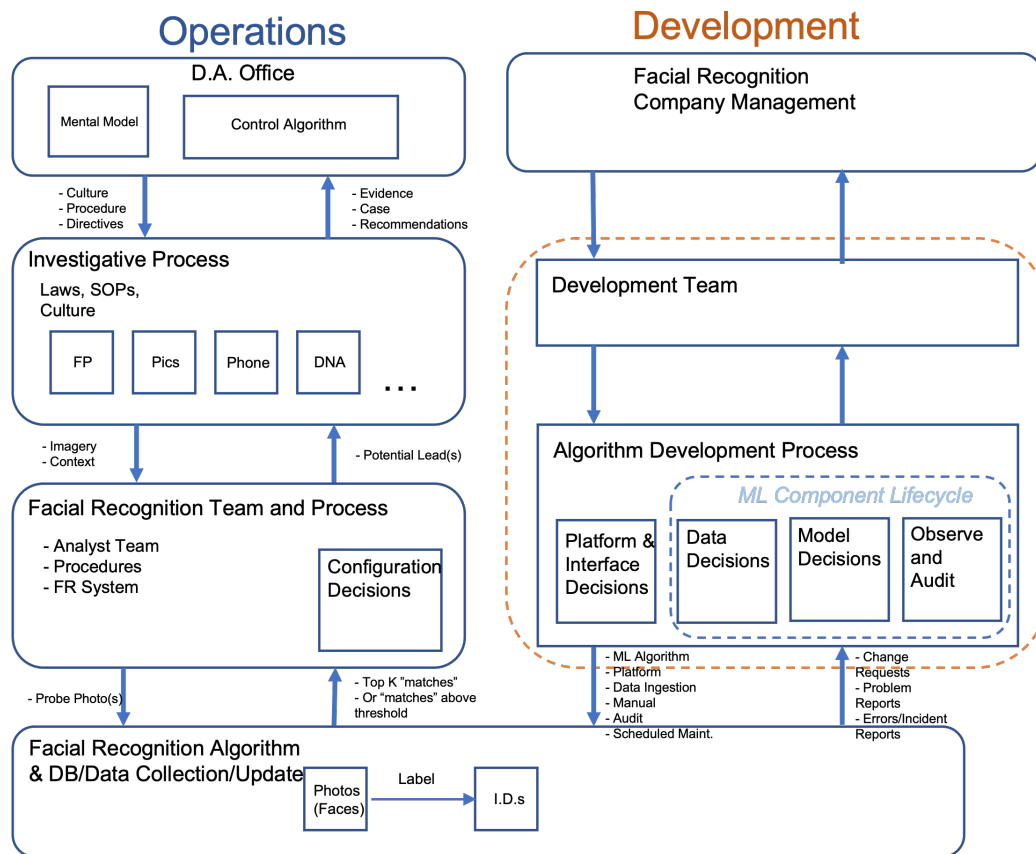


Figure 2.6. Notional CJFR Hierarchical Control Structure

social and ethical impacts downstream.

For instance, one control action in the PFDD control loop is to “define social and ethical success metrics.” *Not providing* success metrics that consider social and ethical impacts is then an unsafe control action which results in hazards at both a high abstraction level (e.g., wrongful warrant issuance or improper investigator confidence in results) and at the lower levels (e.g., imbalance between false match rate and false non-match rate or significant performance difference between demographics).

Once the UCAs are captured in the analysis table and linked to possible hazards, the development team has a structured logical scaffolding with which to work through tensions between low-level development decisions, as well as to reason about balancing traditional performance with social and ethical concerns. In step five, we do this by considering loss

Table 2.9. Facial Recognition Model Decisions Unsafe Control Actions

UCA 2 - Model Decisions				
Control Action	Not Providing	Providing	Too Early/Too Late (Wrong Timing)	Too Low (TL) / Too High (TH)
(1) Architecture: Complexity(Model)				Too complex or too simple of a model [H12,13,14]
(2) Properly Define/Modify Objective Function/Loss and Optimization Procedure	Model behavior and desired behavior mismatch [H6,12,13,14,17]			
(3) Allowed / Recommended Threshold and K determinations				K or threshold is too low [H13] or too high [H12]
(4) Data Augmentation Strategy	Lower-than-possible performance [H12,13,14,17]			
(5) Defense Techniques to Defend against Adversarial attacks (evasion)	Allows for easy creation of effective adversarial examples [H7]	Reduced/alterd model performance [H12,13]		

Table 2.10. Deployment, Observation and Audit Unsafe Control Actions

UCA 3 - Deployment, Observation and Audit Phase				
Control Action	Not Providing	Providing	Too Early/Too Late (Wrong Timing)	Too Low (TL) / Too High (TH)
(1) Establishing a Risk Tolerance for Performance and Social and Ethical Impact Considerations	Does not enable thresholds to trigger meaningful actions to prevent losses [H1,6,7,10,12-22]			Too low a tolerance [H12,13], too high a tolerance (similar to not providing) [H6,7,10,12-22]
(2) Load Testing, Internal Verification and Validation of critical go/No Go Criteria.	Not determining and validating critical go/no go criteria [H1,6,10,12-14]			
(3) Perform Database and Model Update based on requirements defined in Data Decisions.	Stale database of images and associated embeddings [H1,6,11-14,17,22]	New data can lead to undesirable behavior [H15,16,18,19,21]		Too low (not frequently enough) can lead to a stale database and stale embeddings [H1,6,11-14,17,22] Too high(too frequently) can lead to an increased chance of undesirable behavior [H15,16,18,19,21]
(5) Continuous and/or Periodic Monitoring of Social and Ethical Success Metrics (Reference Problem Framing CA2)	If the assumptions of the model and system do not hold, a resulting change in Social and Ethical performance will not be detected and corrected.[H1,6,10,12-14]			Too low, not frequently enough (same as not providing). [H1,6,10,12-14]
(6) Thresholds for alert and/or Corrective Actions	Inaction when the risk limits are surpassed [H1,6,12-14,20]			Too low risks alarm fatigue, too high risks no alarms (same as not providing) [H1,6,10,12-14]

scenarios to identify when unsafe control actions can arise. One potential loss scenario involves the loss of a job by a minority. This might be a result of hazard 1, wrongful warrant

issuance, which might have resulted from false positive matches for all k nearest matches, hazard 12. One photo from the selection of a returned positive matches is used to obtain corroboration by a witness, e.g., through a lineup of photos. Among the possible antecedent control actions, we choose two, *Problem Framing and Data Decisions (PF&DD)* 6—data processing and balancing—and *Model Decisions (MD)* 3—allowed recommended threshold and/or “ k ” determinations. The UCA for PF&DD 6 is that the development team did not sufficiently data balance to tighten the delta in false positive rate between a minority group and the majority. The UCA for MD 3 is that the threshold for the difference in embedding space is too low or in terms of k , that the number selected k is too low—such that with respect to either approach, the correct identity is more often excluded in minority comparisons.

Our step six, *Redesign* can be accomplished in the above example in a number of ways. One approach might be to deal exclusively with data balancing, increasing the relative number of images for the indicated minority that are used to train the encoder model in order to improve that minority’s relative or absolute false positive rate. Another might be to modify or leverage an F_β loss function to accomplish the same ends. Developers might also institute logic based on predetermined NIST-measured deltas between demographics to provide a handicap—i.e., given a delta between demographics and a particular population size allow different k s or thresholds to even effective false positive rates.

Finally, When applying facial recognition systems across large populations, such as the U.S. or even a city such as New York, even a 99.9% accuracy across the population may not be enough. If such a system was, for example, used on each of approximately 5.4 million adults in New York City, 5,400 people, on average, would be misidentified. Likewise, claiming all demographics are above 99% accurate and thus sufficiently equal, also may not be enough to prevent significant harmful disparities. Although these failure modes are well known, STPA offers a repeatable process to identify, discuss, and intervene on problems that arise at the whole-system level, and enables prospective analysis at the design stage expanding on componentwise performance evaluation to create holistic claims of assurance.

2.3 Discussion

Fundamentally, the ideas driving STPA are readily understandable, however, performing conceptual analysis can be challenging for ML-driven systems as these systems do not have

traditional software specifications but rather induced behavior from decisions about data and modeling choices. This is especially true with regard to conducting social and ethical impact analysis where the manifestation of the losses are often spread across a wide variety of stakeholders and difficult to localize to component misbehavior. Thus, the analysis team can tend to spend considerable time working through competing levels of abstraction and multiple variations of STPA losses, hazards, control actions, and feedback loops. Moreover, if the problem is not carefully circumscribed, the team can struggle to understand the right level of “controller” and “controlled process” problems to study. Since the benefit of modeling is often realized in working out unspecified details, these exercises enable the creation of ersatz specifications from which assurance arguments can be derived and paths to defined failures revealed.

A particular aspect of a robust sociotechnical social and ethical impact analysis should also consider whether the resulting product algorithm could be used by the owning company (or sold as a service to another) in a manner not intended in the stated purpose of the system. The reuse or re-purposing of ML-driven sociotechnical systems presents potential challenge for negative social or ethical impact on society (regardless of legality). The analysis should consider these types of uses or re-purposing of ML capabilities for other interests and suggest mitigations or scope statements that warn of these potential issues and the need for separate social and ethical analysis for those uses.

2.3.1 Social and Ethical Analysis Considerations

In our case studies, a number of the controllers or controlled processes involve humans with substantial discretion, such as doctors, pharmacists, and their staffs. Goodhart’s law, “when a measure becomes a target, it ceases to be a good measure,” implies for our analysis that the behavior and effectiveness of controls may change over time [110]. One way that structured analysis such as STPA aids social and ethical risk assessment is by highlighting where this can happen and concretely identifying its consequences. For example, in our PDMP score analysis, such feedback-driven system drift describes undesired behaviors: abandonment by doctors and service refusal by pharmacists, both of which lead to attendant losses. Identification of these behaviors expands the potential to identify hazards and loss scenarios, leading to the identification of new controls to prevent those hazards.

2.3.2 STPA for ML: Analysis Benefits

Conducting an STPA for social and ethical impact forces development teams and other stakeholders to do the necessary work to consider the larger sociotechnical system a model inhabits. Frequently, ML teams are hyper-focused on performance measures and thus miss compositional effects [58]. The STPA approach cultivates a rich appreciation for the complexity ML-driven solutions are joining and provides an effective model with which to reason about harms, as well as the instruments, tools, and methods (both procedural and technical) that we can bring to bear to eliminate or mitigate them.

Another value in using STPA is the mandate to consider carefully the overall goal of the system. In our treatment, this led us to verify that the objective function adopted does not create a mismatch between model behaviors and system goals. For example, applying STPA to the health benefits scoring system studied by Obermeyer [8] could reveal the need for a check for racial and socio-economic disparities resulting from an objective function mismatch early in development and could have enabled a shift to a more appropriate governing optimization. A similar truth may hold for PDMP, more attention should be given to whether the current optimization drives the system in the direction of the overall ostensible goal.

Our research demonstrates the value of abstracting the ML life-cycle as a control structure itself. This new approach recovered potential unsafe control actions during model development in a manner similar to those captured by their analogous operational control loops. The benefits of our approach can be seen in how it breaks the ML life-cycle into manageable phases where each phase has a reasonable number of control actions logically structured to enable responsible development decisions and future audit-ability. Further, this approach allows the model components, control actions, and feedback loops to be decomposed into measurable and verifiable considerations and checks.

Another significant finding is that the STPA process as a whole necessarily provides a traceable path from every derived requirement to contributing unsafe control actions, feedback loops, to the originating stakeholder-defined losses. This property of STPA potentially enables an inherent re-traceability of the reasoning behind decisions in the design, development, and operation of ML systems.

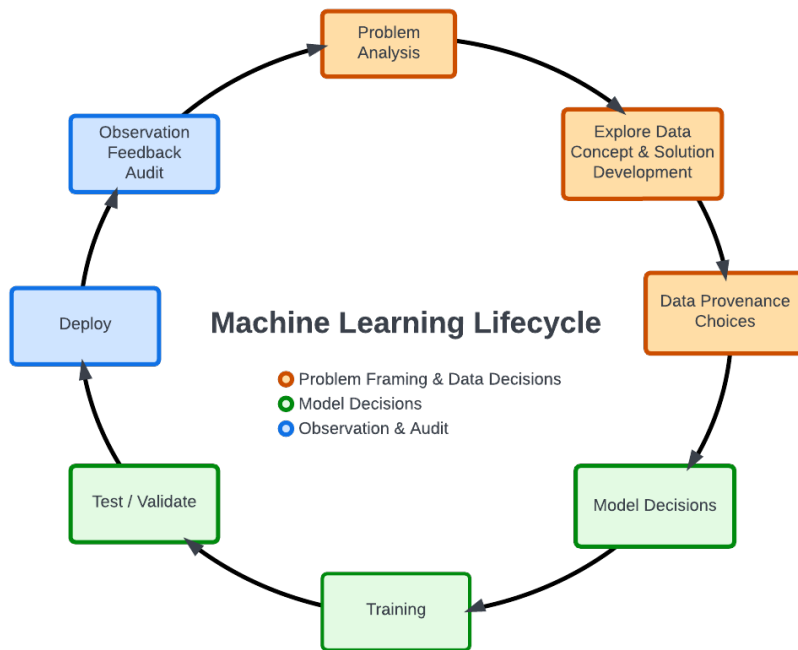


Figure 2.7. ML Life-Cycle

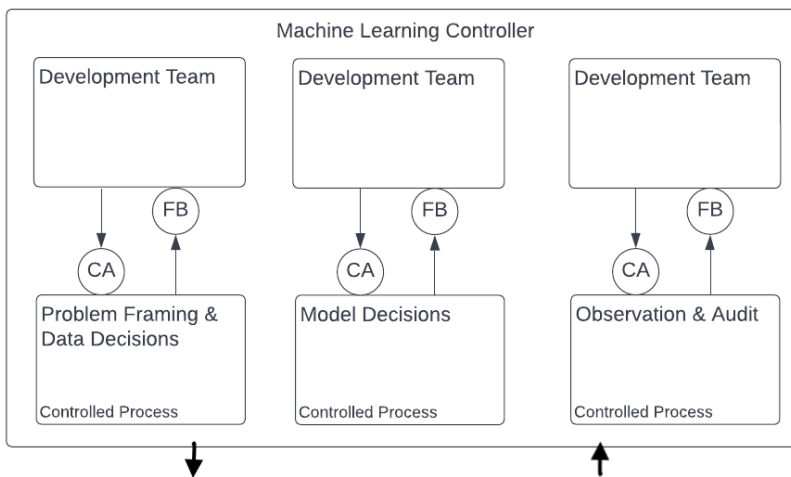


Figure 2.8. Modeling Life-Cycle Control

2.3.3 Considerations For Systems at Scale

An important characteristic of ML systems is that they are applied at scale and encounter populations distinct from those for which they are designed and evaluated. It is necessary to control for this generalization and drift. Practical mechanisms for such control often rely on ongoing measurement of system performance or aggressive retraining, ad hoc solutions to a problem where issues are considered too difficult to foresee. The structured approach of STPA enables the sort of prospective analysis at the control level to define the risks and benefits of various control approaches and avoid the need for unbounded operational decision envelopes.

2.4 Conclusion

This chapter presents the results from case studies which use STPA, a traditional system safety engineering analysis methodology, to assess the social and ethical impact of two very different high-consequence ML-driven systems: 1. a Prescription Drug Monitoring Program *risk score* in the healthcare context and 2. an ML-driven facial recognition system in the criminal justice context. We found that STPA's rigorous approach resulted in a thorough understanding of each subject sociotechnical system and its context as well as the necessary language and reasoning structure with which to build an assurance argument that an overall sociotechnical system is sufficiently trustworthy or safe. The process reveals a trove of hazards and unsafe control actions against which new system requirements for sociotechnical control mechanisms could subsequently be applied to prevent social and ethical losses. These case studies demonstrate that STPA is a promising approach, potentially providing developers with a proven, repeatable and effective process for analyzing potential and existing ML-driven sociotechnical systems for social and ethical impact.

CHAPTER 3: NOTA DNN Defenses and Attacks¹¹

“The most exciting phrase to hear in science, the one that heralds new discoveries, is not ‘Eureka!’ (I found it!) but (instead), ‘That’s funny...’”

—Isaac Asimov

3.1 Introduction

Recent years have seen a steep increase in the number of successful applications of Deep Neural Networks (DNNs) across the sciences, industry, and business. This technology has enabled massive strides forward in areas as disparate as natural language processing [111], astronomy [112], cancer diagnosis [113], protein folding [114], voice identification [115], and machine vision [116], [117]. Despite all of these advances, efforts to leverage DNN technology in safety-critical systems have been severely hampered by the fact that current approaches create models that are fundamentally flawed in that they are highly susceptible to adversarial attacks and related deception techniques. In this chapter we introduce a general defense against evasion attacks motivated by our analysis of the automated classification task and its input space. This first approach we call *naive Adversarial Mixup* (AdvMixup), a new none-of-the-above (NOTA) data augmentation approach for defending deep neural network classifiers from adversarial attacks. We demonstrate how this defense *ostensibly* dispatches state of the art attacks across CIFAR10 and CIFAR100.

Further, we orient NOTA defenses in the literature and discuss why adopting a NOTA paradigm for training DNN classifiers might be a necessary next step toward achieving long-term improved robustness to adversarial evasion attacks. We then explore and discover how NOTA-naive evasion attacks can be trivially defeated by NOTA defenses. We outline several ways in which various mechanisms in each attack can be adjusted to restore effectiveness or at least eliminate trivial failure paths against the broad category of NOTA defenses. We demonstrate these adjustments by modifying eight state of the art attacks that naive

¹¹Portions of this chapter have been previously published by the author as a white paper in collaboration with Professor Armon Barton [29]. Further, research for this chapter was done in collaboration with Professors Barton and McClure.

AdvMixup previously successfully mitigated, then showing how these updated attacks once again produce potent adversarial examples for naive adversarial mix-up defended models. Taking lessons-learned in this process, we modify naive AdvMixup to more effectively seed undefined intermediate portions of input space. We find that this improved strategy for creating NOTA training examples can compensate and restore some, if not most of the lost defense effectiveness in the datasets that we analyze.

3.1.1 Chapter Contributions

In this chapter we introduce a number of offensive and defensive contributions to adversarial evasion attacks in deterministic deep neural networks. They include the following:

- We propose *naive Adversarial Mixup*, a simple technique that produces samples for training a NOTA class to more clearly divide preexisting classes and effectively displace adversarial class partitions in closely adjacent input space, where adversarial examples were previously found.
- We find that naive AdvMix is much more robust than multiple classes of recently proposed defenses when tested against standard attacks, i.e., *Carlini-Wagner L_2 and L_∞* [118], *AutoPGD* [119], *DeepFool* [120], *Square Attack* [121], *Auto Attack* [119], and Jacobian-based Saliency Map Attack [122].
- We evaluate all eight attacks, analyzing their code and how their attack mechanisms encounter NOTA defenses. In this analysis, we identify three inconsequential attack-failure modes which can manifest and by which NOTA defenses can trivially defeat some attacks, they include the following:
 - *Stopping Criteria*: Circumstances under which an attack will presume success and cease perturbation does not preclude NOTA.
 - *Target Selection*: Process used to choose a target non-correct class to perturb an example into does not exclude NOTA as a potential option.
 - *Objective Function*: Objective function does not account for the presence of a NOTA class.
- We adapt each attack to avoid the applicable identified failure modes, while remaining true to each attack’s original methodology.
- We demonstrate that the resulting adaptive attacks largely defeat naive NOTA defenses.
- We adapt our defenses by then modifying the loss function used to create NOTA

augmentation to more evenly distribute NOTA around each class partition within the input space. We also adjust the loss to account for the presence of existing NOTA and seek new regions of input space near the benign example.

- We demonstrate that adapted NOTA defenses result in increased robustness against the NOTA-Aware (NA) attacks while maintaining the same or better robustness against the original versions for each dataset.
- We provide our code for the naive and full AdvMix training algorithm as well as the adapted attacks and fully trained AdvMixup models for CIFAR-10 and CIFAR-100 at (*URL TBD*).

3.2 Background

The DNN defenses we present in this chapter are agnostic of model or datatype, and can be applied to any DNN architecture and in combination with many existing defenses. With respect to evasion attack adaptations, we introduce general concepts that can be applied to any existing attacks to prevent trivial NOTA-paradigm defense successes. Before introducing our defense or attack strategies, it is necessary to discuss, outline, and introduce how the underlying technologies work.

3.2.1 Deep Neural Networks (DNNs)

A *deep* neural network is any neural network made up of many hidden layers of artificial neurons between the input layer and the output layer. The *artificial neuron* was first introduced as a ‘perceptron’ in the 1950s by a psychologist attempting to describe a possible mechanism to bridge the gap between biophysics and psychology [123]. The artificial neuron, like his perceptron, receives inputs from multiple sources—either sensors or other neurons—then aggregates the inputs via a mathematical operation, such as a weighted sum. Finally, it generates an output from that operation that can itself be propagated to the next layer of a network as an input. The general mathematical representation of a single neuron is the expression $\sigma \cdot (w \cdot x + b)$, where σ is a non-linear activation function, w is the *weight* applied to a particular neuron input to the neuron x , and b is the bias, or the negative of the threshold which must be overcome by the inputs for the artificial neuron to activate and provide an output [124].

In general any neural network classifier then can be described as a function $f(x) = y$ where $x \in R^n$ is the input, and $y \in R^m$ is the output¹². The model f also depends on parameters $\theta_i = [W_i, B_i]$ for $1 \leq i \leq k$. The network has k layers, and each layer i within the neural network f has output

$$f_i(x) = \sigma(W_i \cdot x + B_i),$$

where σ is some non-linear activation function, W_i is a weight matrix, and B_i is a bias vector. The number of components in the vectors need not be the same in all the levels. The final output is given by

$$f(x) = f_k(f_{k-1}(\dots(f_1(x)))).$$

In the image classification domain, input x is an $h * w * c$ pixel image such that $x \in R^{hwc}$ and m is the number of object classes for the image. Object classes are denoted by integer codes ranging from 1 to m . The outputs of the function are called *logits*, and the classification of x is given by the equation $y_{\text{pred}} = \text{argmax}(f(x))$.

At this point, it should be mentioned that each layer can be made up of one or more artificial neurons. In generic DNNs, each artificial neuron in a given layer is connected to every artificial neuron in the preceding layer, this is called a *dense layer*, or *fully connected layer*. If a model is comprised entirely of these kinds of layers, it is called a fully connected network. This connectivity means there is no structural enforcement of locality. For instance in computer vision this means that a pixel in the upper right is just as connected to a pixel in the lower left as it is to a pixel that is directly adjacent to it. In these neural networks, data representing an image is said to be “flattened” to one dimension, a process in which a large amount of structural information can be lost [124]. Next, we introduce deep neural networks that have a more limited connection structure that, nevertheless, can capture and encode the information needed for many image classification tasks.

Convolutional Neural Networks (CNN)

As hinted above the structure of visual imagery is encoded in two dimensions by cameras and other sensors. In an image, elements of a picture that are adjacent, above or below one another (spatially near) are often more related than elements that are farther removed or separated.

¹²Where n represents the number of pixels, $n = h * w * c$ (rows, columns and channels), and m represents the number of discrete classes.

This pattern repeats on larger and larger scales. CNNs are an architecture that forces these relationships to be encoded during neural network training. The first convolutional neural networks are attributed to Yan Lecun and Yoshua Bengio [125], largely culminating with their LeNet-5 [126]. In CNNs, kernels or filters are used to combine information from adjacent pixels in an image. For instance a 3x3 pixel kernel has weights for each cell analogous to the individual neuron weight and the kernel has a single bias term. Thus the simple artificial neuron equation $\sigma(w \cdot x + b)$ endures with the kernel now playing the role of w as a matrix of weights.

The kernel enables the model to incorporate data from every pixel that touches the center pixel. It is used to scan across from left to right, and down, row by row, with the kernel centering on every pixel and performing the sum of the element-wise multiplication of the 3x3 kernel weights by the 3x3 neuron inputs and the neuron's bias. In the case of edges where adjacent pixels do not exist, 'padding' pixels can be provided and used. The reduction in connections between layers effected by CNNs results in a significant reduction in computational complexity relative to the number of artificial neuron components. In addition to capturing and encoding new information in a model, CNN layers enabled deeper, and eventually wider neural networks, which are capable of encoding more complex data and relationships.

3.2.2 Deep Learning Vulnerabilities

Broadly speaking, there are three classes of adversarial example attacks defined in the literature: *Exploratory*, *Evasion*, and *Poisoning*. In this chapter, we are primarily concerned with defending against evasion attacks. To provide context, we briefly outline all three. *Exploratory attacks* have to do with those attacks where the adversary does not attempt to achieve misclassification, but instead seeks to gain knowledge of the model by crafting inputs whose resulting output will provide information about the model's internal state, the goal being to reduce uncertainty about a model in support of refining future attacks. *Poisoning attacks* seek to modify the model during training to surreptitiously accomplish some undetected behavior change of the model. Finally, *evasion attacks* imperceptibly modify the artifact or the input to the model to produce a change in classification from the benign (i.e. originally intended) class to some other, untrue class [127]. Defending against this last category is the focus of this and Chapter 4. From this point forward, we refer to these simply

as *adversarial attacks* [128].

Since the initial discovery of adversarial attacks in DNN classifiers in 2013 [34], many dozens of these attacks have been invented, and many more, often narrowly-focused, defenses have been proposed in response. Unfortunately, none of the defenses proposed have thus far substantially and lastingly mitigated the threat posed by these attacks. If their robustness is not largely eliminated in counter testing, then adaptive attacks, or new or even revived attack methods are often brought to bear, which roll back the robustness gains [35]. Some researchers have suggested that this susceptibility may be an unavoidable characteristic of problem formulations in machine learning classification approaches [46]. Others suggest there is a natural trade-off which exists between model accuracy and robustness to adversarial attacks [129]. At this time, no one effective, computationally acceptable, and general approach has been discovered that can shore up DNNs against this and related generalization problems [35], [130].

The general form of evasion attacks is as follows. Adversarial examples are crafted such that given a classifier f , identifying m distinct classes of objects $y_1, y_2, y_3 \dots y_m$, and a benign input x , belonging to the class y_x , an input x' can be crafted such that $f(x') = y_z$ and $y_x \neq y_z$. This is accomplished by adding some perturbation δ_x to x such that $x + \delta_x = x'$, yet x' is still recognized by humans as belonging to the original class y_x [131].

Threat Model

Our threat model assumes the adversary's attack occurs after training and system deployment, (i.e., the adversary cannot manipulate training data). The adversary is assumed to be able to do one of two things. One, they can change the actual artifact in the real world. Examples of this approach would be donning anti-facial-recognition glasses to defeat identification systems, applying an AI-camouflage pattern to a ship or tank to evade wide area motion imagery detection, or simply applying tape to precise positions on a stop sign to fool a self-driving car's image classification system into missing the sign [132]. Two, an adversary with insider access to the data stream can change the direct input to the model by, for instance, altering pixels by imperceptible amounts to fool a classification system into misclassifying the pictured object [118], [131], [133].

We proceed with the following assumptions. An attacker may perturb an image or alter a

physical object to create an adversarial example or representation x' to attack a deep neural net f . Such an attack is considered successful only when the defended network f classifies the resulting image x' as a class other than its benign class y or the NOTA class y_b . In other words, $f(x') \neq y \wedge f(x') \neq y_b$ for a successful attack.

We distinguish between *black-box* and *white-box* attacks against DNNs according to the degree to which an attacker has knowledge of the DNN that they are attacking. A black-box attack is characterized as an attack where the attacker has no knowledge of the DNN's parameters θ , its defenses, or training methodologies. The attacker simply creates adversarial examples using a separately trained deep neural network f' , and transfers and presents those examples to the network f in an attempt to fool it. The white-box attack is strictly more difficult to defend against. It involves an attacker who either has direct access to the original defended neural network or at minimum can create an exact clone, possessing full knowledge of the defended DNN f , the precise training dataset, the parameters θ , the training method, and the defenses used. In this case, the attacker may either create adversarial examples on the original neural network or on a clone f , identical in every respect and parameter to the original.

Projected Gradient Descent(PGD)

Projected gradient descent, is a straight-forward attack which leverages the same optimization that makes DNNs possible in the first place. In PGD, adversarial examples, x' are discovered through gradient ascent and backpropagation [134]. We let $\mathcal{L}(x, y_x)$ represent any loss function whose minimum results in $f(x) = y_x$. Gradient ascent is then employed to iteratively adjust pixels in x such that the loss, $\mathcal{L}$ maximally increases and the resulting image $f(x') \neq y_x$.¹³ For each iterative step, x is adjusted thus:

$$x' = x + \alpha \nabla_x \mathcal{L}(x, y_x),$$

where the gradient vector $\nabla_x \mathcal{L}(x, y_x)$ is the rate of change of the loss, $\mathcal{L}$ with respect to changing x to any non-true class (i.e., not y_x), and α is the learning rate.

Gradient decent is often the basis for learning parameters θ during the training of DNNs. In

¹³This is equivalent to gradient decent on the negative loss.

such training, during each iterative step of gradient decent, θ is adjusted such that

$$\theta' = \theta - \alpha \nabla_{\theta} \mathcal{L}(x, y, \theta).$$

Thus, as shown above, PGD makes use of the same optimization techniques that enabled efficient DNN encoding of the training data in the first place.

AutoPGD (APGD)

Noting that PGD is a frequently used, computationally-cheap method to test adversarial robustness in DNN classifiers, Croce et al. identify two failure modes that can arise and give false assurance of robustness, namely using a fixed step size and second, limiting the loss function to cross entropy (CE) [135]. To improve attack performance, they automate a process to identify a maximally effective step-size for PGD among other variables, as well as provide a new alternative loss function they dub *Difference Logits Ratio (DLR)*. Whereas the loss most often used in PGD is cross entropy, a shift-invariant loss, meaning the order of logits does not alter the output, AutoPGD introduces DLR, which is both shift and *scale invariant*. Scale invariant implies that rescaling will not lead to gradient masking as can occur with CE or Carlini-Wagner losses. DLR is defined thus:

$$\mathcal{L}_{\text{DLR}}(x, y) = -\frac{z_y - \max_{i \neq y} (z_i)}{z_{\pi_1} - z_{\pi_3}},$$

where z is an individual logit from the output of $f(x)$, π is the ordering of the components of z in decreasing order. DLR has been reported to be sometimes better-performing than CE with respect to attack success and more stable than Carlini-Wagner loss(see below) [135].

Carlini and Wagner (CW) Attacks

CW attacks [118] are a suite of attacks that use various optimization methods to find a minimum perturbation adversarial example according to either the L_0 , L_2 , or L_{∞} distance metrics. CW loss takes the form

$$\mathcal{L}_{\text{CW}}(x, y) = -z_y + \max_{i \neq y} (z_i)$$

and each attack is framed as an optimization problem, shown here

$$\begin{aligned} \text{minimize} \quad & D(x, x') + c \cdot \mathcal{L}_{CW}(x') \\ \text{s.t.} \quad & x' \in [0, 1]^n \end{aligned}$$

where the goal is to discover an x' with minimum perturbation distance calculated $D_p(x - x')$ (in some L_p norm) and $f(x') \neq f(x)$. Carlini and Wagner add the additional requirement of the *box constraint* (in order that x' be considered a valid image, all pixels must be in the range $[0, 1]$). As a further requirement, $f(x')$ is minimized, where f is an objective function producing a minimal value when $f(x') = y_t$ – i.e., when the target class is reached and the attack succeeds. The constant c is obtained through binary search and is used to increment or decrement the weight of f . The attack can be targeted where a desired non-true class is provided or untargeted in which case the non-true class with the highest predicted logit z_i output is selected as the target class. Additionally, the attacker may choose a “confidence” level, which controls the margin by which an example target class logit must exceed the true class’s logit.

Deepfool

Researchers Moosavi-Dezfooli, Fawzi and Frossard introduced a novel attack in 2016, called *DeepFool* [120]. The attack is originally conceived and implemented as an L_2 -based attack, but can be adapted to any L_p metric. The DeepFool algorithm is a greedy algorithm that attempts to approximate the minimum distance to the nearest decision boundary then cross it and produce an adversarial example. The approach does not purport to guarantee the smallest possible perturbation, but has been found in practice to yield very small perturbations. The authors believe these perturbations to be good approximations of the minimum. It is not a targeted attack, instead seeking to find the closest region of input space producing a different classification. The pseudocode, shown in algorithm 1 is described by the authors as a gradient descent algorithm using an adaptive step size which it determines at each iteration. It is based on Newton’s iterative method for finding roots of a nonlinear system of equations in the undetermined case.

In simple terms, the algorithm uses Newton’s method to compute an approximation of the vectored minimum distance to the boundary of the *complement* of the input space partition recognized by the model as the correct class, $\hat{r}$. The term w'_y is the difference between the

Algorithm 1: DeepFool: multi-class case [120]

input: Image x , classifier f ;
output: Perturbation $\hat{r}$;
 $y_{\text{pred}}(x) = \text{argmax}(f(x))$;
Initialize $x_0 \leftarrow x, i \leftarrow 0$;
while $y_{\text{pred}}(x_i) = y_{\text{pred}}(x_0)$ **do**
 for $y \neq y_{\text{pred}}(x_0)$ **do**
 $w'_y \leftarrow \nabla f_y(x_i) - \nabla f_{y_{\text{pred}}(x_0)}(x_i)$;
 $f'_y \leftarrow f_y(x_i) - f_{y_{\text{pred}}(x_0)}(x_i)$;
 $\hat{l} \leftarrow \text{argmin}_{y \neq y_{\text{pred}}(x_0)} \frac{|f'_y|}{\|w'_y\|_2}$;
 $r_i \leftarrow \frac{|f'_{\hat{l}}|}{\|w'_{\hat{l}}\|_2^2} w'_{\hat{l}}$;
 $x_{i+1} \leftarrow x_i + r_i$;
 $i \leftarrow i + 1$;
return: $\hat{r} = \sum_i r_i$

gradient with respect to a non-true class and the gradient with respect to the true label. The term f'_y is the difference between the logit corresponding to the class y and the logit corresponding to the true label. The minimum distance $\hat{l}$ is then computed as the result of the argmin of the computed distances. DeepFool assumes a general classifier can be defined as a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^c$ where c is the number of classes and the following mapping of inputs to outputs results:

$$y_{\text{pred}} = \text{argmax}_y f_y(x),$$

where $f_y(x)$ is the output of the function $f(x)$ that is associated with the y^{th} class.

Square Attack

The square attack is a score-based black-box L_2 and L_∞ adversarial attack that does not use local gradient information and thus is immune to gradient masking [121]. It uses a randomized search scheme and perturbations are introduced such that they lie on the boundary of the L_2 -hypersphere or L_∞ -hypercube before their projection back inside the box constraint $([0, 1]^d)$. First, a side-length for the square that will be perturbed is chosen, according to a decreasing schedule. Next, a δ is chosen, if, on applying the δ , the loss

decreases, it is accepted. If not, it is rejected. If the new image classifies in a non-true class, the image is accepted. If not, the algorithm continues, repeating until either successful or the max number of iterations has been completed. The optimization problem this attack seeks a solution for is:

$$\min_{\hat{x} \in [0,1]^d} \mathcal{L}(f(x'), y), \text{ s.t. } \|x' - x\|_p \leq \epsilon, \quad (3.1)$$

where x' is an adversarial example that is created from x , and y is the true label. The rough algorithm for the square attack can be found in Algorithm 2.

Algorithm 2: Square Attack via random search [121]

input: classifier f , point $x \in \mathbb{R}^d$, image size w , number of color channels c , l_p -radius ϵ , label $y \in [1, \dots, K]$, number of iterations N ;
output: approximate minimizer $x' \in \mathbb{R}^d$ of the problem stated in equation 3.1 ;

$x' \leftarrow \text{init}(x)$, $l^* \leftarrow \mathcal{L}(f(x), y)$, $i \leftarrow 1$;
while $i \leq N$ **and** x' is not adversarial **do**
 $h^{(i)} \leftarrow$ side length of the square to modify (according to a schedule);
 $\delta \leftarrow P(\epsilon, h^{(i)}, w, c, x', x)$ (see paper for sampling distributions.);
 $x'_{\text{new}} \leftarrow \text{Project } x' + \delta \text{ onto } \{z \in \mathbb{R}^d : \|z - x\|_p \leq \epsilon\} \cap [0, 1]^d$;
 $l_{\text{new}} \leftarrow \mathcal{L}(f(x'_{\text{new}}), y)$;
 if $l_{\text{new}} < l^*$ **then**
 $x' \leftarrow x'_{\text{new}}$;
 $l^* \leftarrow l_{\text{new}}$;
 $i \leftarrow i + 1$
return: x'

AutoAttack

AutoAttack is considered *the* state-of-the-art adversarial evasion attack. It is a highly effective ensemble of parameter-free attacks, combining the *cross-entropy*-based version of *AutoPGD*, *difference-logits-ratio*-based version of *AutoPGD*, *SquareAttack* and, finally, the *Fast Adaptive Boundary Attack (FAB)* [136]. These separate attacks are used in sequence and until $f_{x_i} \neq f_{x_0}$. The adversarial robustness library (ART) [137], which we use in our testing, substitutes the *DeepFool* attack in for the *FAB* Attack. FAB is distinct in that it extrapolates the hyperplane more precisely than DeepFool. The authors explain, “it would be similar to DeepFool except that our projection operator is exact whereas they project onto the hyperplane and then clip to $[0, 1]^d$.” FAB authors Croce et al, however point out

that a weakness of untargeted FAB is its extensive computational cost as dataset complexity increases and the number of classes increases.

For ease of testing and standardization, the hyperparameters for each attack in AutoAttack are constant across models, datasets, and measurement norms. Given its extensive strength and effectiveness, untargeted AutoAttack is a standardized benchmark used by the adversarial robustness community to compare model defense robustness to adversarial evasion attack in general [138]. The targeted version does not require inputs of targets, but instead attacks using untargeted versions of the subordinate attacks first, then, if unsuccessful in finding an AE for an image, builds off of those untargeted perturbations using targeted versions of the attacks, iterating through the nine most likely incorrect class labels as targets in each targeted attack regime.

Gradient-based Saliency-map Attack

The Gradient-based Saliency-map Attack (GSMA) attack differs from the other attacks mentioned above in that it is exclusively an L_0 attack. This attack is inspired by the Jacobian-based Saliency-map Attack (JSMA) introduced by Papernot et al. [139]. Instead of a Jacobian, however, it uses the gradient, providing knowledge in the form of a saliency map of the most effective pixels to perturb to achieve the desired results of a change in class while also minimizing the total perturbation. Algorithm 3 provides the pseudo-code for the attack [139].

Algorithm 3: Gradient-based Saliency Map Attack [139]: x is the benign image, y^* is the target network output, f is the function learned by the network during training. Υ is the maximum distortion, and θ is the change made to pixels.

```

input :  $x, y^*, f, \Upsilon, \theta$ 
 $x' \leftarrow x$ ;
 $\Gamma = 1 \dots |x|$   $\triangle$  search domain is all pixels;
while  $F(x') \neq y^*$  and  $\|\delta_x\| < \Upsilon$  do
     $S = \text{saliency\_map}(\nabla f(x'), \Gamma, y^*)$ ;
    Modify  $x'_{i_{\max}}$  by  $\theta$  s.t.  $i_{\max} = \text{argmax}_i S(x, y^*)[i]$ ;
     $\delta_x \leftarrow x' - x$ ;
return  $x'$ 

```

3.2.3 DNN Defenses

Given the extensive literature that has accumulated regarding new methods to find effective, fast and cheap evasion attacks, significant effort has been expended seeking effective measures to mitigate or eliminate these threats. Here we describe some common and often used defenses.

Adversarial Training Data Augmentation

The best known method for increasing DNN robustness remains Adversarial Training [?], [131]. This is a data augmentation technique which adds adversarial examples to the true label class. The most common and effective form uses PGD which in many cases leads to significantly improved adversarial robustness.

TRADES

Researchers Zhang et al. introduce *TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES)* [129]. TRADES is centered around the concept that there is necessarily a trade-off between natural error (1 - accuracy) and what they describe as robustness error (R_{rob}). Their approach to defending neural networks is reflected in the following minimization:

$$\min_f \mathbb{E} \left\{ \underbrace{\mathcal{L}(f(\mathbf{x}), y)}_{\text{for accuracy}} + \underbrace{\max_{\mathbf{x}' \in \mathbb{B}(\mathbf{x}, \epsilon)} \mathcal{L}(f(\mathbf{x}), f(\mathbf{x}')) / \lambda}_{\text{regularization for robustness}} \right\}$$

This expression reflects the trade-off intended in the TRADES defense between the natural and robust errors. The first term in expression 3.2.3 provides loss between the predicted class and the true label, thereby driving accuracy to increase. The last term uses PGD to encourage the output to be *smooth* by maximizing the loss provided between a natural example $\mathbf{X}$ and a corresponding adversarial example $\mathbf{X}'$ divided by the term λ , the regularization parameter. This has the result of pushing the decision boundary of the classifier away from data samples. As λ decreases the natural accuracy decreases and the robust accuracy increases. The authors adjust this parameter for a given dataset to find the desired trade-off between model accuracy and robustness to adversarial examples.

3.2.4 NOTA Defenses

Our hypothesis is that DNN model robustness, indeed, resistance to the broad class of adversarial example attacks can be improved with approaches that defend models by introducing a ‘None-of-the-Above,’ or NOTA, class during training. The premise is that with carefully crafted NOTA augmentation examples, we can continuously ‘seed’ the benign data-point-sparse space between data-dense regions of a classifier’s input space throughout training and leverage the DNN to classify this vast space as NOTA. This approach is suggested as a potential remedy to the ‘inevitability’ of adversarial examples under the current predominant training paradigm by Shafahi et al. [21]. While covering all of the data-sparse regions of input space would be impractical, we hypothesize it might be sufficient to cover the regions that define the boundaries between and adjacent to classes and let the DNN generalize the remaining input space as NOTA. Further, we hypothesize that we can define these boundary regions and construct a NOTA class by generating and training on samples related to the original training data.

Shafahi et al. showed mathematically using isoperimetric inequalities that if the partition of each class corresponds to a partition that takes up less than half the input space, then by their proofs, *every* example in *any* class will necessarily be within an extremely small L_p distance from an adversarial example [21]. We conceive then, that one goal of a NOTA approach would be to coax the DNN model into designating more than half of the input space as NOTA. This then, should bring every example in any other class into a, according to Shafahi et al., calculably very close proximity of the NOTA-class partition [21]. Though, the result cannot be said to rule out the possibility of adversarial examples from other classes, it, at a minimum, begins to crowd out the space of opportunity for other classes, especially in the presence of NOTA examples specifically chosen to surround the true label class partition tightly. With this theoretical basis, we seek to empirically explore this NOTA-based approach as an effective defense.

This training strategy is agnostic of the attack method or the neural network structure. It changes the structure of how the input space is partitioned by the classifier through training with an additional NOTA class. The approach is modular and able to be added to any DNN classifier, and may augment nearly any present or future defense strategy. The entire approach has a moderate, but fixed computational cost which only occurs during training.

3.2.5 Previous NOTA Approaches

Barton [140] first demonstrated NOTA approaches to adversarial robustness and deep neural network defense with “Boundary Padding”, a data augmentation approach which sought to insert a "padding class" between various classes in input space by linearly interpolating between two images and subsequently labeling the resulting example as a padding class (i.e., NOTA). Results were very effective against most attacks at the time.

3.3 Naive Adversarial Mixup

Naive AdvMix was conceived after preceding research showed that various methods of creating NOTA class examples—such as using linear interpolation in the input space, or mixing methods in latent lower dimensional space using auto-encoders—showed promise at defending DNNs against adversarial examples produced by the CW attack suite using confidences of 20 or higher [29], [140]. However these methods struggled to perform against low-confidence Carlini Wagner L_p attacks. This same work showed improvement in robustness resulted from including PGD examples in the NOTA class [29].

3.3.1 Methodology

An important additional inspiration resulted from Zhang et al’s *mixup* [141] algorithm for instantiating linear behavior between training examples and increasing regularization and resistance to adversarial attack. Though we discard the label-mixing aspect, we use the simple mixing expression, $\lambda \cdot X_1 + (1 - \lambda) \cdot X_2$ for the image data as well as randomization of λ in a new way to create our NOTA class examples. We conceive this NOTA augmentation approach, Adversarial Mixup, as a way to more closely surround correctly labeled regions of input space with NOTA class examples. However, instead of ‘mixing’ two separate training examples, as well as their labels, and training as that label combination, as is done in mixup, AdvMix ‘mixes’ a single training set example x_1 with an adversarial example, x'_1 , derived using the PGD attack. The resulting AdvMix image is labeled as the NOTA class and added to training on the fly. Note that *no mixing of the labels occurs*, all produced examples are instead labeled as the NOTA class and trained as such.

In naive AdvMix two variations of NOTA are produced on the fly and added to the training batch before batch-training commences, *mean AdvMix* and *uniform AdvMix*. In mean

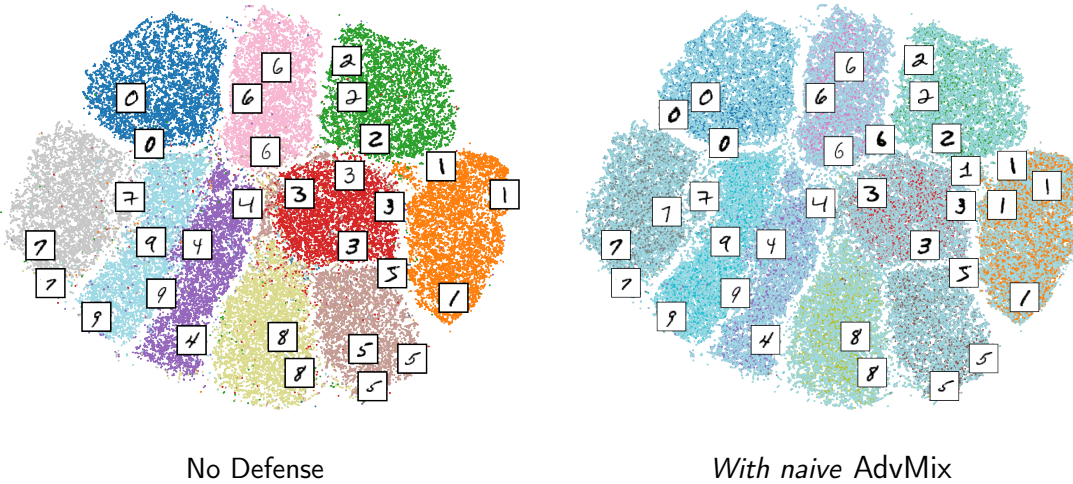


Figure 3.1. tSNE Plots of MNIST: Light-Blue Dots on Right, Surrounding Each Class, Are AdvMix Datapoints.

AdvMix, λ is set to 0.5 and limited gaussian noise is added to the resulting image. In uniform AdvMix, λ is set to a random number between 0.05 and 0.95 and a weighted average of the benign and adversarial example is performed. The resulting NOTA to benign data ratio is two to one, making it by far the largest class in the dataset. This large representation of NOTA vs any other class reflects the intuition that regardless of the number of finite classes that are defined, the vast majority of possible inputs in the input space do not correspond to any of the specified classes and, thus, should instead be assigned to the NOTA class.

3.3.2 Experimental Setup

We use standard Wide Residual Networks (WRN) with dropout and batch normalization, configured as suggested by Zagoruyko et al. as the base model architecture for our experiments [142]. Dropout is set to 30% drop probability during training. Across all models for CIFAR-10 and CIFAR-100, we use a wide resnet, 12 units deep and 6 units wide, i.e., WRN-12-6, a common setup for these datasets. For each model we use an *ADAM* (adaptive moment estimation) optimizer with default settings, $b_1 = 0.9$ and $b_2 = 0.999$ [143] with sharpness-aware minimization (SAM) [144]. ADAM assists the model in efficiently converging on a solution, whereas SAM, by seeking out minima of the training loss landscape

and minimizing loss curvature, smooths the boundaries between partitions expressed in input space.

When performing mini-batch gradient descent training, two separate batches of 32 are drawn from the training set for each training cycle. The first uses standard dataset augmentation, a 10% shift up or down, and 10% shift left or right, as well as horizontal flipping and up to a 15 degree rotation clockwise or counter clockwise. The second batch is a ‘clean’ batch drawn from a separate iterator without data augmentation. Each batch is used both as a benign training batch and also to create AdvMix NOTA examples based on each training example. Our datasets are split before training such that 4% of the former training set are quarantined as a validation set to enable early-stopping model selection, based on a combination of best validation accuracy and best validation adversarial robustness. Every 150 batches during training, the same 30 examples are used (previously separated from the validation set) to create 30 untargeted, zero-confidence CWL_2 adversarial examples with a maximum of 10 iterations. These adversarial examples are then used to calculate the training ‘attack success rate’ (ASR), which is used in model selection. ASR is calculated by determining the number of adversarial examples that successfully drove the model to classify the image as a class *other* than the NOTA class or the true label. The number of successful examples divided by the total number provided results in the ASR.

Table 3.1. Model Accuracies

CIFAR-10	
Model	Accuracy
Undefended WRN-12-6	92.31%
Naive AdvMix WRN-12-6	90.93%
CIFAR-100	
Undefended WRN-12-6	70.53%
Naive AdvMix WRN-12-6	68.13%

The test set is strictly reserved for testing a specific model that has been previously selected using ASR and accuracy performance on the validation set. In testing, accuracy is calculated from the full test set. ASR is calculated on the preselected model using adversarial examples created on the first 100 test set samples. We test against Carlini Wagner L_2 , L_∞ , AutoPGD-CE, AutoPGD-DLR, DeepFool, Square Attack, AutoAttack and GSMA. In parameterized attacks

we set max iterations to 100, i.e., CW L_p attacks, square attack and DeepFool (L_2). Both AutoPGD attacks, Square attack, and AutoAttack are performed in both L_2 with max epsilon of 0.5 (maximum distance between x and x' by specified L_p metric) and L_∞ with max epsilon of 0.031, the distances specified for each by Robust Bench [145]. DeepFool is tested with both max $\epsilon = 0.1$ and $\epsilon = 0.5$. Finally GSMA is tested with a maximum single pixel perturbation of 0.031 and gamma settings of $\gamma = 10\%$ and $\gamma = 20\%$ (the maximum percentage of pixels in the image which may be perturbed).

3.3.3 Results

Table 3.2. Model Attack Success Rate vs. SOTA Attacks (100 Max Iterations)

Naive AdvMix Model vs. Naive SOTA Attacks														
Models	C&W Suite		APGD CE		APGD DLR		Square Attk		DeepFool		AutoAttk, Untrgtd		GSMA	
<i>norm, max ϵ or confidence</i>	L_2 , Conf:0	L_∞ , Conf:0	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.1	L_2 , ϵ :0.5	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	$\gamma = 0.1$	$\gamma = 0.2$
CIFAR-10														
No Defense	99%	100%	100%	100%	87%	99%	10%	67%	99%	99%	100%	100%	96%	100%
Naive AdvMix	15%	1%	6%	6%	43%	54%	13%	62%	10%	9%	6%	6%	91%	99%
CIFAR-100														
No Defense	99%	100%	100%	100%	100%	100%	39%	86%	71%	71%	100%	100%	91%	99%
Naive AdvMix	44%	0%	30%	30%	72%	69%	46%	33%	25%	23%	30%	30%	31%	82%

As recorded in Table 3.2, naive AdvMix results in noteworthy performance across a broad cross-section of attacks. According to current standings as of writing, Robust Bench records the best CIFAR-10 L_2 and L_∞ performances against untargeted AutoAttack as having robust accuracies of 84.86% and 70.69%, equating to ASRs of 15.14% and 29.31% respectively [138]. This version of AdvMix scores an ostensibly astounding L_2 and L_∞ ASR of 6% and 6% for CIFAR-10 on these attacks. Further, the top performing CIFAR-100 defense against L_∞ AutoAttack on Robust Bench has 42.67% robust accuracy, equating to 57.33% ASR. AdvMix comes in with an ASR of just 30%, once again far and away an apparent improvement. With respect to other main-stream attacks, we, additionally, show that AdvMix records noteworthy reductions to ASR for most. There are two exceptions, however, L_∞ Square Attack, which remains a high 60s ASR, and GSMA. GSMA maintains its high effectiveness in CIFAR-10 even for a relatively small gamma of 10%. Even so, this would seem an impressive advance in favor of increased adversarial robustness in classification systems.

However, there are some curiosities in table 3.2. One is that Square Attack remains so effective against AdvMixup and, yet, AutoAttack, which leverages square attack as one of its components, does not register an ASR above the model’s natural error rate. One key to this mystery is that NOTA augmentation defense approaches in general rest on a fundamental change to the existing training paradigm, (i.e., the addition of the NOTA class and the assessment that examples which are predicted in that class are adversarial examples). With this information in mind, we can look closely at the attacks and analyze how the mechanisms within the attack respond to this change. Carlini et al. point out that most DNN defense developers fail to adequately consider how their defense could be successfully attacked or circumvented [146]. By adapting the mechanisms we identify here, we should be able to get substantially closer to a more accurate assessment of the adversarial robustness of these models and future models of this genre.

3.4 NOTA-Aware Attacks

Recognizing that NOTA augmentation changes the input landscape as described above, it becomes apparent that key mechanisms in adversarial evasion attack code must be adjusted to become more explicit in their handling of classes other than the true label. At a minimum, mechanisms that involve the *stopping criteria*, *target selection* and *objective function* must be reevaluated in light of this change.

Stopping Criteria

Stopping criteria are the predetermined circumstances under which attack code will presume success and cease exploring the input space by continuing to modify the adversarial example. It is no longer the case under the NOTA paradigm that just any class other than the true class will do. In fact, the most abundant class, the one that is most likely to represent the closest decision boundary or next highest probability or logit, is likely to no longer correspond to a successful attack, as the goal of NOTA approaches is that these criteria will result in NOTA. For instance, some untargeted attacks set one condition for success to be to change the predicted class such that $\text{argmax}(F(x')) \neq y_{True}$. Properly restated in a NOTA paradigm this should be changed to the following:

$$(\text{argmax}(F(x')) \neq y_{NOTA}) \wedge (\text{argmax}(F(x')) \neq y_{True})$$

This prevents attacks from immediately stopping when NOTA is the predicted class.

Target Selection

In many attacks, if a particular target class is not provided (i.e., a desired false class to perturb the image into), the attack chooses a ‘best,’ random, or otherwise determined class that is not the true label. If present, such a mechanism must be altered to additionally exclude the NOTA class as a target. This is true in CW attacks [147], DeepFool [120], as well as targeted AutoPGD and targeted AutoAttack [135].

Objective Function

In some cases, an objective function which does not account for the presence of a NOTA class can result in driving perturbations directly into the NOTA class, rather than through it or away from it. As the non-true class with the closest data points to every training example, NOTA frequently has the steepest gradient away from the correct class and can act as a local minimum, or honeypot. This may require adjustment to the objective function. This is true in CW attacks and is worth exploring in gradient descent-based attacks.

3.4.1 Methodology

In this section, we modify existing state of the art attacks as coded in the Adversarial Robustness Toolbox [137] to eliminate the trivial NOTA failure paths identified above.¹⁴ After discussing the necessary changes in each attack we demonstrate their improved effectiveness by attacking the same AdvMix model trained above.

NOTA-Aware Carlini and Wagner L_∞ and L_2

For a description of Carlini and Wagner’s attacks see Section 3.2.2 [147]. Each attack requires an adjustment regarding selection of a target class. In the loss function for both attacks, two changes were necessary. First, when the variable z_{other} is calculated, the NOTA class must be eliminated from consideration, since this will be the target class for the attack. Second, the attack has not succeeded until z_{other} is greater than the highest of either the logit of the true label class or the logit of the NOTA class plus the input confidence score. The L_∞

¹⁴In addition to their inclusion in supplementary material, before publication we will make these modified NOTA attacks available for public use on Github.

attack required an additional change. The generate function is altered to eliminate NOTA as an acceptable potential solution when minimizing distance from the benign sample.

NOTA-Aware Deep Fool

DeepFool is described in detail in Section 3.2.2 [148]. This attack required both an adjustment regarding choice of targeted class and an adjustment to the stopping criteria. We first removed NOTA as an option when the algorithm ranks potential targets in its selection process. We also ensure that the attack will not stop in the case where NOTA is the predicted class after perturbation, unless it has reached maximum iterations without finding a successful adversarial example.

NOTA-Aware Square Attack

In Andriushchenko et al.'s Square Attack [121], we first evaluated the stopping criteria. We altered the criteria so that the attack was only successful if it resulted in a classification other than the benign class *and the NOTA class*. We also prevented the NOTA class from being chosen as the initially correct label, in the event the true labels are not provided. We investigated changes to the loss function to incorporate a term that incentivized increased loss between the adversarial example and the NOTA class in addition to the benign class, but this resulted in less effective attacks against NOTA defenses. We believe this was due to the fact that the loss was used as a litmus test to determine if a change was an improvement or not and the two separate terms rarely reinforced and, more often, destructively interfered with one another resulting in no applied alterations. The strongest attack resulted from eliminating NOTA as a legitimate stopping class in the stopping criteria and eliminating it as a possible benign class.

NOTA-Aware Auto Projected Gradient Decent (APGD) CE and DLR

Both the cross entropy (CE) and difference logits ratio (DLR) versions of Croce et al.'s original Auto PGD attack are described in section 3.2.2 [135]. In both versions of the APGD attack it is necessary to ensure that if the labels are not provided, that the model does not predict the NOTA class as the benign example's class in creating the labels. More importantly, we also alter the generate function such that it will not stop the attack when the NOTA class is predicted, but instead keep iterating forward. For the DLR version, we ensure that NOTA

is not designated as any of the three logits that are used to calculate the difference logits ratio loss.

For the CE version of APGD it was worth investigating an adjustment to the loss function considering the introduction of a massive class like NOTA. We adjusted the loss to maximize a combination of the benign label class loss and the NOTA label class loss by taking their mean. We investigated various combinations of the two to determine the most effective, from solely maximizing the loss between the adversarial example and the NOTA class to solely maximizing the loss between the adversarial example and the original class label (i.e., the original attack loss). These various loss combinations had little effect on attack efficacy, so we retained the original loss as a result.

NOTA-Aware AutoAttack

As AutoAttack [136] leverages four separate parameter-free subordinate attacks, as described in Section 3.2.2, it is necessary to ensure that the NOTA-aware version of AutoAttack calls the NOTA-aware versions of each of these subordinate attacks. Additional changes are required as well. The check in the function `_run_attack` that is performed after each subordinate attack is complete regarding whether the resulting adversarially-perturbed example is misclassified and thus a successful adversarial example is updated to exclude NOTA in addition to the true label. Finally, the section of code that deals with the targeted attack requires updating as well. This attack generates a set of targets to iterate through to find an adversarial example. Originally this set included all non-true label classes. We remove NOTA from the set of potential targets.

NOTA-Aware Gradient-based Saliency Map Attack (GSMA)

In GSMA the only change we found necessary was to remove the NOTA class from a generated list of potential random targets. One further avenue for future investigation would be to determine to what degree the gradient used in the saliency calculation could be refactored to exclude the portion of the gradient resulting from NOTA. This has not been tried as yet, but is an avenue for further research for potential improved efficacy against NOTA defenses.

3.4.2 Experimental Setup

Our experimental setup remains much as it was established in Section 3.3.2. We retest the naive AdvMix model with the same hyper-parameter settings as before and compare the results to both the undefended and naive AdvMix models' performance against the original attacks.

3.4.3 NOTA-Aware Attack Results

Table 3.3. Naive AdvMix vs. NOTA Aware Updated SOTA Attacks

Naive AdvMix Model Vs NOTA Aware SOTA Attacks														
Models	C&W Suite		APGD CE		APGD DLR		Square Attk		DeepFool		AutoAttk, Untrgt'd		GSMA, $\theta = 0.031$	
<i>norm, max ϵ or confidence</i>	L_2 , Conf:0	L_∞ , Conf:0	L_2 , $\epsilon:0.5$	L_∞ , $\epsilon:0.031$	L_2 , $\epsilon:0.5$	L_∞ , $\epsilon:0.031$	L_2 , $\epsilon:0.5$	L_∞ , $\epsilon:0.031$	L_2 , $\epsilon:0.1$	L_∞ , $\epsilon:0.5$	L_2 , $\epsilon:0.5$	L_∞ , $\epsilon:0.031$	$\gamma = 0.1$	$\gamma = 0.2$
CIFAR-10														
No Defense v orig attk	99%	100%	100%	100%	87%	99%	10%	67%	99%	99%	100%	100%	96%	100%
Naive AdvMix v orig attk	15%	1%	6%	6%	43%	54%	13%	62%	10%	9%	6%	6%	91%	99%
Naive AdvMix v NOTA attk	99%	6%	8%	6%	94%	98%	13%	62%	18%	13%	98%	100%	75%	98%
CIFAR-100														
No Defense	99%	100%	100%	100%	100%	100%	39%	86%	71%	71%	100%	100%	91%	99%
Naive AdvMix v orig attk	44%	0%	30%	30%	72%	69%	46%	33%	25%	23%	30%	30%	31%	82%
Naive AdvMix v NOTA Attk	49%	30%	30%	30%	99%	98%	43%	47%	35%	27%	98%	98%	40%	93%

As is evident in Table 3.3, adapting the state-of-the-art adversarial evasion attacks to the NOTA paradigm resulted in some massive changes in performance for naive AdvMix. In some cases, such as APGD-CE, CW L_∞ and DeepFool, significant robustness is retained, even if at a slightly higher ASR. The remaining attacks, however, are largely devastating to naive AdvMix, this is especially true for APGD-DLR, AutoAttack and CW L_2 adapted attacks. Although even the original GSMA is very effective against the CIFAR-10 dataset, it is less effective against the AdvMix-defended CIFAR-100 dataset. In CIFAR-100, the NOTA-aware version of the GSMA attack results in an increase in the attack's effectiveness with respect to that more complex dataset.

The story is much the same between the two datasets, although naive AdvMix appeared to defeat most of these state-of-the-art attacks, adapting the attacks to the change in training paradigm that NOTA represents, by altering the stopping criteria, target selection, and/or

objective functions resulted in overcoming the AdvMix defense scheme, as presented thus far. However, this performance *does* suggest that there is *something* to NOTA augmentation approaches, even if attack success is largely restored in this instance—some attacks remain largely defeated, such as CW L_∞ and DeepFool. In the next section, we take steps to improve the NOTA defense.

3.5 Adaptive NOTA Defenses

A number of questions arise from the preceding section regarding naive AdvMix’s performance. What can we learn from the attacks where effectiveness against the defense is restored? Specifically, can anything be inferred about the partitioning of input space that results from AdvMix? One conclusion we draw from the initial performance against the original attacks is that AdvMix clearly establishes a NOTA presence near and around the true class partition. However, as shown in the adapted attacks, this presence is not eliminating or fully replacing regions where other classes reside near the true label partition. Equally important is to analyze the attack adaptations for which the defense largely retains its adversarial robustness even after adaptation.

It seems clear from APGD-CE and CW L_∞ performances that the naive AdvMix defense is strong with respect to loss gradients that result from cross entropy. Additionally, the largely retained robustness to DeepFool that increases in robustness with increased distance seems to imply a significant stand-off distance between data dense partitions of input space. What seems to endure as weaknesses of the defense are the attacks that leverage the DLR loss, APGD-DLR and AutoAttack, and attacks that do not use loss gradients at all, such as Square Attack and GSMA, especially under the L_∞ norm.

3.5.1 Methodology

With the above lessons learned, we set to the task of identifying techniques that can be brought to bear to increase NOTA presence and crowd out the model’s partitions for other classes in data-sparse regions surrounding data-dense regions. Considering the geometric arguments of Shafahi et al. [21], one might seek to cause the model to increase the ratio of input space allotted to NOTA. One method to accomplish this is to increase the ratio of NOTA to benign samples that the models see during training. Additionally, it appears from

the previous experiments that AdvMix can benefit from using a different loss with PGD to explore the input space and plant NOTA examples in as-yet largely unreached regions. We have been using PGD with CE exclusively to create AdvMix examples; could the approach improve significantly by using the DLR loss to create NOTA as well, given the defense's poor performance against attacks that leverage that loss?

Another concern is with using any optimization scheme that seeks solely to maximize loss from the true label class. As NOTA regions are planted and reinforced over the course of training epochs, given present loss formulations, it is likely that NOTA will be planted in nearly the same input space locations epoch after epoch. Each new NOTA example then results in a steeper gradient to that portion of the input space, which in turn, will result in higher likelihood that future NOTA will be created in close to the same place. This can have a reinforcing effect creating an analogue of a gravity-well effect, thus, clumping NOTA regions near but not surrounding a class's partition space. Instead, the desired behavior is to surround homogeneously-classed datapoint-dense regions of input space with NOTA, while also populating data-sparse regions between them with NOTA. One method to prevent such clumping and spread NOTA more evenly in the input space would be to add a term to the PGD loss function that also maximizes the loss between the perturbed example and the NOTA label. Alternatively, some NOTA examples could be crafted by perturbation in the direction maximizing loss with NOTA exclusively and other NOTA examples created by maximizing loss with the true label class. Finally, one last mechanism aimed at reaching new regions in input space is inspired by targeted AutoAttack. The targeted version of AutoAttack iterates through all of the classes, setting each as a target class with which to minimize loss at each perturbation. A brute force approach would overwhelm the model with NOTA examples, however one could achieve a desired ratio by randomly selecting the target from all classes save NOTA and the benign class for each sample. With these potential mechanisms in mind, we investigate two new strategies to improve AdvMix's robustness performance across the full cross section of NOTA-aware state-of-the-art attacks.

Simple Reflexive NOTA-Aware (SRNA) Adversarial Mixup

The SRNA defense retains all previous characteristics of AdvMix, except it switches between two different losses. One PGD loss maximizes the cross entropy with respect to the true label class as in naive AdvMix. The second loss maximizes the cross entropy loss with respect to

Table 3.4. Model Accuracies

CIFAR-10	
Model	Accuracy
Undefended WRN-12-6	92.31%
Naive AdvMix WRN-12-6	90.93%
SRNA AdvMix WRN-12-6	91.48%
MRNA AdvMix WRN-12-6	91.04%
CIFAR-100	
Undefended WRN-12-6	70.53%
Naive AdvMix WRN-12-6	68.13%
SRNA AdvMix WRN-12-6	69.11%
MRNA AdvMix WRN-12-6	66.84%

the NOTA class. This strategy is elegant in its simplicity, alternating between pushing away from the true label class to plant NOTA, and pushing away from existing NOTA to plant NOTA where it does not already exist.

$$\mathcal{L}(y_{\text{true}}, y_{\text{pred}}) = \begin{cases} \text{CE}(y_{\text{true}}, y_{\text{pred}}) & \text{when } \beta \leq 0.5 \\ \text{CE}(y_{\text{nota}}, y_{\text{pred}}) & \text{otherwise} \end{cases}$$

where $\beta \sim U(0, 1)$.

Multi-modal Reflexive NOTA-Aware (MRNA) Adversarial Mixup

The MRNA defense retains many previous characteristics of AdvMix, but there are a number of loss function modes used and the ratio of NOTA examples to benign examples is six to one vice the previous two to one in naive AdvMix and SRNA. The first mode of loss used to create NOTA is **Adaptive Cross Entropy Loss** NOTA. These examples are crafted from PGD using one of the two following compound loss expressions for each adversarial example at 50% probability:

$$\mathcal{L}(y_{\text{true}}, y_{\text{pred}}) = \begin{cases} \alpha \cdot \text{CE}(y_{\text{true}}, y_{\text{pred}}) + (1 - \alpha) \cdot \text{CE}(y_{\text{nota}}, y_{\text{pred}}) & \text{when } \beta \leq 0.5 \\ \alpha \cdot \text{CE}(y_{\text{true}}, y_{\text{pred}}) - (1 - \alpha) \cdot \text{CE}(y_{\text{nota}}, y_{\text{pred}}) & \text{otherwise} \end{cases}$$

where $\beta \sim U(0, 1)$, $\alpha \sim U(0.25, 0.75)$, and CE is the standard cross entropy loss. The second mode of loss used to create NOTA is **Adaptive Difference Logits Ratio Loss** NOTA. These

examples are created from PGD using one of the two following compound loss expressions for each adversarial example at 50% probability.

$$\mathcal{L}(y_{\text{true}}, y_{\text{pred}}) = \begin{cases} \alpha \cdot \text{DLR}(y_{\text{true}}, y_{\text{pred}}) + (1 - \alpha) \cdot \text{TDLR}_{\text{nota}}(y_{\text{true}}, y_{\text{pred}}) & \text{when } \beta \leq 0.5 \\ \alpha \cdot \text{DLR}(y_{\text{true}}, y_{\text{pred}}) - (1 - \alpha) \cdot \text{TDLR}_{\text{nota}}(y_{\text{true}}, y_{\text{pred}}) & \text{otherwise} \end{cases}$$

where $\beta \sim U(0, 1)$, $\alpha \sim U(0.25, 0.75)$, DLR is the difference logits ratio loss from AutoPGD and $\text{TDLR}_{\text{NOTA}}$ is the targeted difference logits ratio from AutoPGD, hard-coded to always target NOTA. Both APGD losses have two cases, which result in two types of adversarial examples. Both cases have a first component expression that maximizes the loss between the example and the true class. The second component either maximizes or minimizes the loss between the example and the NOTA class. We believe the components that point back toward NOTA act to stabilize training ASRs by strengthening existing NOTA partitions, whereas the components that point away from NOTA help to establish new regions of input space as NOTA.

The remaining two sources of NOTA are inspired by targeted AutoAttack. These modes begin with adversarial examples created as a byproduct of both the Adaptive Cross Entropy NOTA and the Adaptive Difference Logits Ratio NOTA. The first is *Targeted Cross Entropy*, taking mixed-source adversarial examples as input, it randomly selects non-true target labels for each example and then uses PGD to perturb the example further toward the target class by minimizing the cross entropy loss to that target.

$$\mathcal{L} = -1 \cdot \text{CE}(y_{\text{targets}}, y_{\text{pred}})$$

The final source of NOTA, *Targeted Difference Logits Ratio* also takes mixed-source adversarial examples as input, randomly selects non-true labels as targets and then uses PGD to perturb the example further toward the target class by minimizing the difference logits ratio loss to that target represented in the following.

$$\mathcal{L} = -1 \cdot \text{DLR}(y_{\text{true}}, y_{\text{pred}}, y_{\text{targets}})$$

TDLR is the same loss function described by Croce et al. in “Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks” [135] and in section 3.2.2. These targeted modes of NOTA are intended to plant NOTA deeper within data-sparse regions and on maximum gradient paths between all classes.

3.5.2 Experimental Setup

Our experimental setup remains much as it was established in Section 3.3.2. A single difference is the increase in NOTA to benign data ratio to six to one in MRNA. This increase allows us to accomplish two aims. First, it allows us to create a uniform and mean AdvMix NOTA version of each one of the above described loss-function modes. Second, we test whether an increase in the ratio of NOTA to benign class assists in achieving better bias toward NOTA outside data-dense regions. We test SRNA AdvMix and MRNA AdvMix with both the original attacks and the NOTA-aware attacks. Again, max iterations are set to 100 for square attack, Carlini Wagner attacks and DeepFool. We collect data for an L_2 bound of $\epsilon = 0.5$ and an L_∞ bound of $\epsilon = 0.031$.

3.5.3 Adaptive NOTA Defense Results

Table 3.5 shows the performance of all models tested in the chapter (these are all deterministic models).

CIFAR-10

With respect to the CIFAR-10 dataset, our modified AdvMix models SRNA and MRNA recover considerable ground in defending against the newly introduced NOTA-aware attacks. Among NOTA-Aware Carlini Wagner Attacks, L_2 0 Confidence is reduced from a staggering 99% ASR against naive AdvMix to 20% ASR against MRNA. Likewise the modified AdvMix models both showed significant improvement over GSMA $\gamma = 10\%$ with ASRs falling from 75% (naive AdvMix) to 13% (SRNA) and 24% (MRNA) as well as a 40% reduction against $\gamma = 20\%$. Both MRNA and SRNA versions of AdvMix show improvement against NOTA-Aware AutoPGD-DLR over naive AdvMix's 94% ASR, but even so, they remain quite high at 54% ASR (MRNA) and 55% ASR (SRNA) for the L_2 norm at $\epsilon = 0.5$. Finally, our modified AdvMix models both show improvement against untargeted NOTA-Aware AutoAttack in both norms, with a more significant reduction by MRNA AdvMix in both norms with an L_2 of 52% ASR and L_∞ of 89%, down from naive AdvMix ASRs of 98% and 100% respectively.

Table 3.5. Full Table of Deterministic Models vs. NOTA-Aware Updated SOTA Attacks

Full Table of Deterministic Models Vs NOTA Aware SOTA Attacks														
Models	C&W Suite		APGD CE		APGD DLR		Square Attk		DeepFool		AutoAttk, Untrgtd		GSMA	
<i>norm, max ϵ or confidence</i>	L_2 , Conf:0	L_∞ , Conf:0	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	L_2 , ϵ :0.1	L_2 , ϵ :0.5	L_2 , ϵ :0.5	L_∞ , ϵ :0.031	$\gamma = 0.1$	$\gamma = 0.2$
CIFAR-10														
No Defense	99%	100%	100%	100%	87%	99%	10%	67%	99%	99%	100%	100%	96%	100%
Orig AdvMix v old attk	15%	1%	6%	6%	43%	54%	13%	62%	10%	9%	6%	6%	91%	99%
SRNA AdvMix v old Attk	17%	0%	9%	9%	18%	46%	9%	46%	11%	10%	9%	9%	7%	60%
MRNA AdvMix v old Attk	22%	0%	11%	11%	11%	22%	11%	57%	16%	10%	11%	11%	26%	58%
Orig AdvMix v NOTA Attk	99%	6%	8%	6%	94%	98%	13%	62%	18%	13%	98%	100%	75%	98%
SRNA AdvMix v NOTA Attk	39%	9%	9%	10%	55%	91%	9%	44%	14%	11%	60%	96%	13%	51%
MRNA AdvMix v NOTA Attk	20%	21%	11%	11%	54%	76%	11%	61%	16%	10%	52%	89%	24%	59%
CIFAR-100														
No Defense	99%	100%	100%	100%	100%	100%	39%	86%	71%	71%	100%	100%	91%	99%
Orig AdvMix v old attk	44%	0%	30%	30%	72%	69%	46%	33%	25%	23%	30%	30%	31%	82%
SRNA AdvMix v old Attk	34%	0%	26%	26%	51%	65%	26%	65%	16%	14%	26%	26%	7%	26%
MRNA AdvMix v old Attk	41%	0%	30%	30%	44%	43%	37%	64%	14%	12%	30%	30%	20%	29%
Orig AdvMix v NOTA Attk	49%	30%	30%	30%	99%	98%	43%	47%	35%	27%	98%	98%	40%	93%
SRNA AdvMix v NOTA Attk	99%	26%	26%	26%	85%	95%	28%	61%	28%	22%	83%	98%	3%	28%
MRNA AdvMix v NOTA Attk	43%	30%	30%	30%	86%	92%	36%	65%	20%	16%	81%	94%	4%	27%

CIFAR-100

In the CIFAR-100 dataset the story is again much the same with some interesting additional insights provided by observations in a more complex dataset. With the CIFAR-100 findings it is important to first note that the natural error rate in this dataset is already 30% to 33% depending on the model.¹⁵ In this dataset SRNA fell apart against NOTA-Aware CWL_2 , returning a 99% ASR, whereas MRNA returned 43% ASR, only 10% more than its natural error. All AdvMix CIFAR-100 models returned NOTA-Aware CWL_∞ numbers at or even

¹⁵See Table 3.4 for specific accuracies.

better than their natural error rate. All three models, again performed very well against the NOTA-Aware cross-entropy version of AutoPGD likewise with ASRs at or *better* than their natural error rates across both norms. MRNA AdvMix performs best against NOTA-Aware DeepFool with L_2 $\epsilon = 0.1$ returning a 20% ASR (13% below the natural error rate) and L_2 $\epsilon = 0.5$ returning a 16% ASR (17% below the natural error rate). Contrary to the moderate defense performance in CIFAR-10, both MRNA and SRNA AdvMix eliminate the GSMA problem at both $\gamma = 10\%$ and $\gamma = 20\%$, against both NOTA-Aware and original attacks with all ASRs well below the natural error rate for each model. Overall performance however is not an unqualified success, significant challenges remain against AutoPGD-DLR and AutoAttack and even Square attack in the L_∞ norm. Against NOTA-Aware difference logits ratio AutoPGD, the new models do show improvement over naive AdvMix’s L_2 ASR of 99% and L_∞ ASR of 98%, but only very modestly. MRNA AdvMix performs the best overall with an anemic L_2 ASR of 86% and L_∞ ASR 92%. Against AutoAttack, the story is similar with MRNA giving the best results, improving over naive AdvMix’s ASRs of L_2 : 98% and L_∞ : 98% to MRNA results of L_2 : 81% and L_∞ : 94%.

3.6 Discussion

In this chapter we introduced a new NOTA defense, naive AdvMix and demonstrated its apparent defeat of a wide swath of state of the art attacks including bringing untargeted AutoAttack ASR down to at or below natural error rate in both CIFAR-10 and CIFAR-100 models. We then analyze the eight very different state-of-the-art attacks and discover three general trivial failure modes NOTA defenses may trigger in evasion attacks to defeat them, *stopping criteria*, *target selection* and *objective function*. Each of these are often generalized in attack mechanisms such that a none-of-the-above (NOTA) class can easily act as a honeypot and serve as the easiest change of class. However, we demonstrate that attacks that are NOTA-adapted using these three mechanisms are effective against naive AdvMix. Taking lessons learned about the input space from the original attacks, naive AdvMix, as well as the change in performance of the adapted attacks, we propose and test two very different approaches to modifying the AdvMix defense, Simple Reflexive NOTA-Aware Adversarial Mixup and Multi-modal Reflexive NOTA-Aware Adversarial Mixup. We then retest and show that these defenses show improvement against all attacks, completely defeating some. Some, nevertheless, remain largely potent. Clearly, despite the

improvements demonstrated with MRNA and SRNA AdvMix, the NOTA-Aware AutoAttack and NOTA-Aware AutoPGD-DLR results, in particular, show there remains work to be done to improve robustness to acceptable levels. In the next chapter, we investigate how combining the NOTA approaches introduced here with known Bayesian concepts and methodologies as well as a novel distribution and sampling approach may close much of this remaining gap.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:

Improving NOTA Robustness with Bayesian Neural Networks¹⁶

In the previous chapter, we discussed the concept of the *none-of-the-above* (NOTA) DNN defense. We also introduced a new NOTA defense approach, as well as a general methodology for adapting existing evasion attacks to improve their success against NOTA defenses. Finally, we took lessons learned from the successes of the adapted attacks and offered improved NOTA defenses. In this chapter, we again take aim at improving adversarial robustness, now combining the NOTA defense from chapter 3 with Bayesian stochastic methods. We hypothesize that NOTA and Bayesian methods will interact and reinforce each other. Just as before, leaving attacks at status quo without adapting them to defeat the defense that is applied is not acceptable. Therefore, we again update all state-of-the-art attacks previously adjusted in Chapter 3 by adding state-of-the-art modifications for stochastic defenses, in addition to the NOTA adaptations. These modifications have been shown to overcome previous stochastic model defenses [149], [150].

Research has shown that stochastic ensembles of models can be both more accurate and more robust to adversarial attacks [151]. Pinot et al. demonstrate, through mathematical proofs, a guaranteed lower bound on a model's robust accuracy given specified levels of injected noise sampled from an exponential family distribution [152], [153]. However, the levels of noise needed to guarantee strong adversarial robustness in the standard training paradigm can, and do, lead to significant reductions in accuracy in these implementations [152]. Research by Athalye et al. shows that these defenses, when near-state-of-the-art accuracy is maintained, are susceptible to attacks modified to use sampling within the attack to produce an *expectation over transformation* (EOT) [149]. In an addendum white paper, Dhillon and Carlini point out a flaw in the Athalye analysis and implementation of a dropout-based defense, fixing it and identifying an additional, and even more effective, attack against some stochastic defenses, i.e., the *deterministic whitebox transfer (DWBT) method* [150]. We update each of the eight attacks from the previous chapter and implement both of

¹⁶Research for this chapter was done in collaboration with Professor Patrick McClure.

these techniques in each attack. Finally, we introduce a new Bayesian DNN technique, produced using exponential distribution family noise. This technique demonstrates a synergy, capturing defensive strengths observed from both the Bernoulli distribution and the Gaussian distribution in contributing to adversarial robustness.

Chapter 4 Contributions

In this chapter we make the following contributions:

- Novel combination of existing stochastic approaches with underlying NOTA-defended base models for increased adversarial robustness at near state-of-the-art accuracy.
 - Monte-Carlo Dropout (MCD) BNN with underlying NOTA-defended base model
 - LA BNN with underlying NOTA-defended base model
 - Novel combination of Laplace with Dropout (LAD), with underlying NOTA defended base model.
- Implement DWBT attacks for each of the eight NOTA attacks from Chapter 3.
- Implement EOT adaptations for each attack from Chapter 3.
- Demonstrate robustness gains by both NOTA MCD BNN and NOTA LA BNN defenses.
- Demonstrate complementary effects of the novel combination of Laplace with MC Dropout NOTA defense resulting in SOTA accuracy and robustness, exceeding Robust Bench performances of the latest top defenses by wide margins.

4.1 Background

A number of statistical tools are used in the research covered in this Chapter. The most central to our discussion being Bayes Theorem. Bayes theorem enables a statistician to update predicted probabilities of an event based on new information. The theorem is stated thus mathematically:

$$\begin{aligned}
 p(\theta|D) &= \frac{\text{likelihood} * \text{prior}}{\text{evidence}} \\
 &= \frac{p(D|\theta) * p(\theta)}{p(D)}
 \end{aligned}
 \tag{4.1}$$

The equation can be read as the posterior probability $p(\theta|D)$ is equal to the probability of the data D , given prior knowledge of the event θ , multiplied by the prior probability of the event, $p(\theta)$, divided by the probability $p(D)$ (the evidence). Unfortunately, a well-known central problem in Bayesian statistics is the intractability of calculating the denominator, $p(D)$. In many cases, there is no known closed-form solution to $p(D) = \int p(D|\theta) * p(\theta) d\theta$ [154].

To translate this into terms useful for our coming discussion, the posterior $p(\theta|D)$, in the case of ML, represents a *distribution* of acceptable parameter configurations (denoted by θ , and meaning the weights and biases of each artificial neuron in the DNN) from which one can sample to produce a single particular model, with models that accurately represent the training data being more probable to sample. The right hand side of the equation provides a path to mathematically derive such a distribution. In the following sections, we will discover the ways in which we can use this to our advantage to produce accurate and robust models.

4.1.1 Maximum A Posteriori (MAP) Estimation

Maximum a posteriori (MAP) estimation is a method for finding the most likely model based on Bayes theorem. The MAP estimate represents a single-point estimate of the largely unobservable distribution of parameters, given the available empirical data [155]. Mathematically, the MAP estimation θ is defined by:

$$\begin{aligned}\theta_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}} [p(\theta|D)] \\ &= \underset{\theta}{\operatorname{argmax}} \left[\frac{p(D|\theta)p(\theta)}{p(D)} \right]\end{aligned}\tag{4.2}$$

As stated before, however, the evidence, $p(D)$, the term in the denominator, is inaccessible. Nevertheless, as the evidence itself is not dependent on θ , we can ignore the denominator and maximize the numerator for the solution. Additionally, as a logarithm is a monotonically increasing function ¹⁷, we can take the logarithm of the numerator above to simplify the problem and get:

¹⁷Monotonically increasing—the significance of which is, the logarithm of a function reaches its maximum value at the same points as the function it acts on.

$$\begin{aligned}
\theta_{\text{MAP}} &= \underset{\theta}{\operatorname{argmax}} [\log(p(\theta|D)p(\theta))] \\
&= \underset{\theta}{\operatorname{argmax}} [\log(p(\theta|D)) + \log(p(\theta))]
\end{aligned} \tag{4.3}$$

Finally, maximizing the posterior $p(\theta|D)$, as we've done above, is equivalent to minimizing a *loss* [156].

$$\theta_{\text{MAP}} = \underset{\theta}{\operatorname{argmin}} [-\log(p(D|\theta)) - \log(p(\theta))] \tag{4.4}$$

For classification DNNs, the likelihood term, $-\log(p(D|\theta))$, equates to cross-entropy and the prior term, $-\log(p(\theta))$, equates to L_2 regularization when a mean-zero isotropic Gaussian prior is used [157].

4.1.2 Bayesian Model Averaging

Thus, MAP estimation corresponds to the most ubiquitous method of training deterministic DNN models [157]. However, the MAP is a point estimate of an entire distribution of models, which can potentially be leveraged. We will cover how we can recover that distribution of models from the MAP in future sections.

We previously touched on research that shows that ensembles of models can improve accuracy [158]. Pinot et al. demonstrate that *Bayesian* ensembles can also provide provable lower bounds on adversarial robustness [152]. *Bayesian model averaging* provides a technique for leveraging the ability to utilize an ensemble of models, each sampled from $p(\theta|D)$, or an approximation of $p(\theta|D)$. As discussed in Chapter 3, in deterministic models the goal is to learn to predict the label y for a given input x using information learned from a training dataset, D_{Train} . In supervised ML, D_{Train} is a set of label-input pairs, $[x_i, y_i]$. In the probabilistic paradigm, this changes to learning the probability of y_{Test} given x_{Test} and D_{Train} , i.e., $p(y_{\text{Test}}|x_{\text{Test}}, D_{\text{Train}})$. In ML, we use models with parameters, θ . A model then corresponds to computing $p(y_{\text{Test}}|x_{\text{Test}}, \theta_{\text{Train}}, D_{\text{Train}})$.

We now derive how we can use this concept to leverage a posterior, or a posterior approxi-

mation by using marginalization and the definition of conditional probability:

$$\begin{aligned}
p(y_{\text{Test}}|x_{\text{Test}}, D_{\text{Train}}) &= \int p(y_{\text{Test}}, \theta|x_{\text{Test}}, D_{\text{Train}})d\theta \\
&= \int p(y_{\text{Test}}|x_{\text{Test}}, \theta, D_{\text{Train}})p(\theta|x_{\text{Test}}, D_{\text{Train}})d\theta \\
&= \int p(y_{\text{Test}}|x_{\text{Test}}, \theta)p(\theta|D_{\text{Train}})d\theta
\end{aligned} \tag{4.5}$$

Above, the first step uses the definition of marginalization. The second step uses the definition of conditional probability. The final step accounts for the fact that the parameters, θ encode the information that the dataset, D_{Train} contains regarding the relationship between an input and its paired label. It is important to note, however, that this integral is often intractable, given parameters often number in the millions and are real numbers. Therefore, some additional approximation is necessary. We cannot test with the whole distribution, since the above integral is intractable. However, once we have an approximation of the distribution to draw our parameters, θ , from, we can use *Monte Carlo Sampling* to estimate the expectation through sampling:

$$p(y_{\text{Test}}|x_{\text{Test}}, D_{\text{Train}}) \approx \frac{1}{k} \sum_{i=1}^k p(y_{\text{Test}}|x_{\text{Test}}, \theta^i) \tag{4.6}$$

where $\theta^i \sim p(\theta|D_{\text{Train}})$ and k is the number of samples taken.

4.1.3 Laplace Approximation (LA)

In this section, we introduce the *Laplace Approximation* (LA), the first of two methods we will leverage to access stochastic ensembles of different models. The LA is a local Gaussian approximation of the posterior, $p(\theta|D)$, (our distribution of acceptable models) *centered* around the MAP estimate. LA builds on the intuition that a set of models with analogous behavior can be found in a Gaussian distribution around the MAP estimate for the model parameters. Further, this Gaussian distribution is shaped by a *Hessian*, or second-order derivative, of the loss function for the parameters [159]. Whereas a first-order derivative provides the ‘slope’ of change for a parameter, the second order derivative provides the ‘curvature’. The results of the calculated Hessian are used to provide a local approximation of the posterior distribution for each parameter, θ . Hence, high curvature means small changes

in a parameter have significant effects on the loss and, thus, a tighter distribution about the MAP for such a parameter would be necessary. Whereas, low curvature means large changes in a parameter can result in smaller effects on loss, thus a broader distribution of values could be used on the parameter in those cases.

Mathematically, the LA uses a Hessian of the loss $\mathcal{L}(D, \theta)$ around θ_{MAP} to formulate a Gaussian approximation of the posterior, $p(\theta|D)$, per [159]:

$$p(\theta|D) \approx \mathcal{N}(\theta, \theta_{\text{MAP}}, \Sigma) \quad (4.7)$$

where $\Sigma := (\nabla_{\theta}^2 \mathcal{L}(D, \theta)|_{\theta_{\text{MAP}}})^{-1}$,

The Hessian of the loss is calculated as follows [159]:

$$\nabla_{\theta}^2 \mathcal{L}(D; \theta)|_{\theta_{\text{MAP}}} = \underbrace{-\gamma^{-2} I}_{\text{log-prior Hessian}} - \underbrace{\sum_{n=1}^N \nabla_{\theta}^2 \log p(y_n | f_{\theta}(x_n))}_{\text{log-likelihood Hessian}}|_{\theta_{\text{MAP}}}, \quad (4.8)$$

where γ is weight decay and I is the identity matrix. We use a diagonal Hessian approximation, as commonly done in the literature [159], [160]. Combining the two equations above provides an approximation of the posterior distribution from which one can repeatedly sample to provide model parameters to construct an ensemble of models. We will use the resulting ensemble of models and Bayesian model averaging as tools to help us build robust and accurate classification models.

4.1.4 Monte Carlo Dropout (MCD)

The second method we will leverage to query an ensemble of models is *Monte Carlo Dropout* (MCD). Srivastava et al. introduce *Dropout* as a regularizing technique to prevent neural networks from over-fitting training datasets [161]. Dropout involves randomly ‘dropping’ a portion of artificial neurons in a neural network during its training. The effect is that training occurs using an exponentially large distribution of different ‘thinner’ neural network models to update weights in training. However, as originally implemented, at test time, the full neural network is instead implemented, but with proportionately reduced weights for the neurons to compensate for the increased number in this full model. The original authors suggest and perform some Bayesian model averaging using Monte Carlo sampling to provide

comparisons to their final full implementation. Gal et al., however, do extensive testing, demonstrating Monte Carlo dropout in DNNs and the advantages these implementations can provide [162]. In MCD, sampling is performed both during training and during testing, in order to access the exponential number of different neural networks it creates. In, MCD training and testing, each neuron i in the hidden layers is associated with a Bernoulli random variable b_i . Thus, our simple neuron function from chapter three now becomes:

$$b_i \sigma \left(\left[\sum_{j=0}^{d-1} w_{i,j} x_{i,j} \right] + w_{i,d} \right), \quad (4.9)$$

where $b_i \sim \text{Bern}(1 - p_{\text{drop}})$ and p_{drop} is the dropout rate. With Monte Carlo dropout, our MAP estimate, equation 4.4 still applies with appropriate sampling. Clearly the negative log prior term does not change with sampling as it is the prior distribution. Likewise, the evidence term does not change either. However, the log likelihood term does change under dropout. The change can be shown mathematically by:

$$\begin{aligned} p(D|\theta) &= \sum_b p(D, b|\theta), && \text{marginalization} \\ &= \sum_b p(D|b, \theta) p(b|\theta), && \text{conditional probability} \\ &= \sum_b p(D|b, \theta) p(b). && \theta \text{ does not change with } b \end{aligned} \quad (4.10)$$

Finally, we can arrive at the following approximation using Monte Carlo sampling:

$$p(D|\theta) \approx \frac{1}{m} \sum_{j=1}^m p(D|b^j, \theta), \quad (4.11)$$

where $b^j \sim p(b)$, is a particular sampled neural network dropout configuration, $p(b) = \prod_i p(b_i)$, and m is the number of samples taken.

4.1.5 Adversarial Strategies Against Stochastic Defenses

We leverage the strongest subset of the NOTA-aware attacks introduced in Chapter 3, excluding only those which were reduced to at or below natural error in the deterministic models. Please refer to Section 3.2.2 for a description of the original attacks and Section

3.4 for an explanation of how each was adapted to overcome NOTA-defended models. With respect to stochastic attacks, our review of the literature revealed two main strategies to overcome stochastic defense methods, which can each be applied to any existing attacks.

Expectation Over Transformation (EOT)

In addressing how to defeat the stochastic defenses, *Stochastic Activation Pruning (SAP)* [163] and *Randomization* [164], Athalye et al., propose altering attacks such that they calculate the *expectation* of the gradients over many samples vice taking a single gradient sample to direct each perturbation [149]. The expectation of a model's predictions can also be used in these types of attacks, as in EOT Square Attack.

Deterministic White Box Transfer Attacks (DWBT)

Finding that the SAP defense, when correctly implemented, actually largely successfully defended against the EOT modifications, Dhillon et al. show that adversarial examples built with the undefended white box version of a model and then transferred to the defended model were subsequently 70% successful against the SAP defense [150]. This approach is especially effective against defenses that essentially mask gradients.

EOT and DWBT represent sound repeatable strategies to alter existing attacks to make them effective against stochastic defenses. Additionally, each has been validated against existing stochastic defenses. We take each of these successful attack strategies and build them into the state-of-the-art NOTA-aware attacks we adapted in Chapter 3 to test our Bayesian ensembles of NOTA models.

4.2 Methodology

In Chapter 3, we demonstrated the high accuracy and increased robustness of NOTA-defended deterministic models. Research by Pinot et al. demonstrates increased robustness results from stochastic BNNs, however at a cost of a significant reduction in accuracy [152]. Here we investigate whether combining a NOTA-defended deterministic model with two separate approaches to BNNs, LA and MCD, can result in near state-of-the-art accuracy along *with* state-of-the-art adversarial robustness. We also investigate whether these two methods of BNN implementation, MCD and LA, may be combined together to synergistic effect, in a

novel Bayesian distribution sampling technique, LA with Dropout (LAD). We choose these approaches, LA, MCD, and LAD specifically because they can easily be applied, in many cases, to many pre-trained deterministic models, as we will show below.

4.2.1 Laplace Approximation with Dropout (LAD)

Pinot et al. proved that injecting noise sampled from the exponential family of distributions can guarantee a provable lower bound to robustness [153]. In LAD, we sequentially inject noise from two exponential family distributions, the Gaussian distribution in the case of LA and the Bernouli distribution in the case of MCD. In LAD, a local Gaussian approximation of the posterior is learned using a dropout-enabled neural network. Below, we derive the mathematical underpinnings for combining LA and MCD.

We decided to investigate whether this combined distribution might provide synergy in defending NOTA DNN models, when we observed complementary defense performance when comparing NOTA with MCD and NOTA with Laplace (LAP). While each showed significant and mostly state of the art robustness against all, effectively 16 attacks ¹⁸, NOTA with LAP tended to perform better (i.e., below natural error) against EOT attacks, whereas MCD tended to perform better against DWBT attacks. A natural question then is, if combined do the strengths or the weaknesses confer to the joint distribution?

Bayesian Model Averaging for LAD

In LAD, we can accomplish Bayesian model averaging by substituting the approximated MCD likelihood, derived in equation 4.11, into the Bayesian model averaging approximation, equation 4.6. The result is:

$$\begin{aligned} p(y_{\text{Test}}|x_{\text{Test}}, D_{\text{Train}}) &\approx \frac{1}{k} \sum_{i=1}^k p(y_{\text{Test}}|x_{\text{Test}}, \theta^i), \\ &\approx \frac{1}{k} \frac{1}{m} \sum_{i=1}^k \sum_{j=1}^m p(y_{\text{Test}}|x_{\text{Test}}, b^j, \theta^i), \end{aligned} \tag{4.12}$$

¹⁸We implemented each of the eight state of the art NOTA-aware attacks from Chapter 3 as both a DWBT attack and a EOT attack, therefore, 16 attacks.

where θ^i is sampled from our LA of the posterior ($p(\theta|D)$), $b^j \sim p(b)$, k is the number of samples from our LA, and m is the number of MCD samples taken for each LA sample.

Computing Likelihood Hessian for LAD

Combining Laplace Approximation with Monte Carlo Dropout additionally requires re-working the computation of the likelihood term in the Hessian to properly include dropout sampling. First, we can derive how this is approximated in the Laplace Approximation using sampling of the training dataset:

$$\begin{aligned} & \nabla^2 - \log(p(D|\theta)), \\ & \nabla^2 n \cdot \frac{1}{n} \sum_{l=1}^n -\log(p(D_l|\theta)), \\ & \nabla^2 n \cdot \mathbb{E}_{p_{\text{Train}}(D)} \left[-\log(p(D_l|\theta)) \right], \\ & \approx n \cdot \frac{1}{s} \sum_{l=1}^s \nabla^2 - \log(p(D_l|\theta)), \end{aligned} \tag{4.13}$$

where $D_l \sim p(D_{\text{Train}})$, s is the number of samples, n is the number of datapoints in the training set, and $p_{\text{Train}}(D)$ is the empirical distribution for the training data. Finally, we can substitute our approximation of likelihood with MCD, equation 4.11, into this approximation of the Hessian, equation 4.13, to gain our likelihood Hessian for LAD:

$$\approx n \cdot \frac{1}{s} \sum_{l=1}^s \nabla^2 - \log \left(\frac{1}{m} \sum_{j=1}^m p(D_l|b^j, \theta) \right), \tag{4.14}$$

where $D_l \sim p(D_{\text{Train}})$, s is the number of LA samples, n is the number of datapoints in the training set, $b^j \sim p(b)$, and m is the number of MCD samples for each LA sample. This completes our derivation for LAD. In Section 4.3, we will evaluate and compare the performance of this technique combined with a NOTA-defended MAP estimate, against separate NOTA defended MCD and LA models.

4.2.2 Bayesian Ensembles of NOTA Models

In the classical training paradigm for DNN classifiers, each deterministic model has its own partitioning of input space. In attacking these models, adversarial examples can be

found using this single deterministic partition. However, although different deterministic models trained on the same dataset are likely to agree on the partitioning of data-dense regions of input space, this is not necessarily so regarding the data-sparse regions of input space. Under this classical paradigm then, in ensembles, an attacker must find an adversarial example that fools the ensemble of models. However, if large portions of the input space correspond to inputs that, properly, should not be associated with any class, yet are, adversarial examples are often still easily found. Shafahi et al. show that in this paradigm, adversarial examples necessarily exist [21]. Ensembles created through Bayesian approaches alone have demonstrated increased robustness, however, state-of-the-art robustness and near state-of-the-art accuracy have not been simultaneously achieved, particularly when tested with EOT and/or DWBT attacks.

Under the NOTA paradigm of DNN training from Chapter 3, our deterministic NOTA defenses attempt to combat adversarial examples by surrounding the data-dense regions of input space for each benign class with tightly-drawn NOTA class examples. Thus, the intended effect is to pull a NOTA decision boundary tightly around each class's data-dense partition of the model's input space. Further, regions outside of these regularized, training-data-dense partitions, should be largely recognized as the NOTA class. There are a number of ways to do this in the deterministic model, we adopt two, namely: one, increase the ratio of NOTA training data to benign training data, to effectively train the model to bias itself toward a NOTA classification, and two, additionally seed the space between and around training-data-dense regions with targeted PGD examples using both CE and DLR losses.

Combining Bayesian ensembles with a NOTA-defended base model, we believe will reinforce the tight NOTA decision boundaries around data-dense regions of input space, recall figure 3.1. Our intuition is that each model drawn from its distribution source will model data-sparse regions differently. However, each sampled model will have largely partitioned the data-sparse regions in favor of the NOTA class, as it is the largest class by at least an order of magnitude and likely the nearest training example, given the nature of AdvMixup NOTA generation. Our intuition is that the only class the resulting sampled models will agree on in the data-sparse regions should be the NOTA class.

4.3 Experiments

For ease of comparison, we approach our investigation into combining NOTA with MCD, LAP, and LAD in a parallel manner with our work from Chapter 3 on deterministic models. To save computational resources, we have eliminated attacks which were reduced to at or below natural accuracy in the deterministic defense, such as AutoPGD-CE [135], Carlini and Wagner L_∞ [147] and DeepFool [120], and, instead, have focused on attacks that were still highly successful in challenging the NOTA defense, such as AutoPGD-DLR [135], Carlini and Wagner L_2 [147], AutoAttack with both L_2 and L_∞ bounds [135], and Square Attack [121] at the L_∞ bound.

4.3.1 Experimental Setup

We use the models trained for the experiments performed in Chapter 3, therefore the experimental setup detailed in Sections 3.3.2, 3.4.2, and 3.5.2 still applies with the following additions and caveats. In some circumstances, models were left training longer after model selection for Chapter 3 and at times achieved a (what was thought to be) better combination of validation accuracy and validation robustness. In those situations, we have chosen the new model as a base model or MAP estimate for this chapter, vice the Chapter 3 model. In some circumstances, this actually resulted in worse deterministic performance on the test set (opting for the higher accuracy with a slightly lower adversarial robustness as a trade-off does not pay off here deterministically). We provide the updated deterministic performance in these situations for comparison in table 4.1. Even in such circumstances, we consider this to be an even better testing opportunity, to see how these combinations of approaches can interact synergistically.

Hyper-parameter selection for LA and LAD consists first of fitting the LA model to the benign Training dataset, which involves calculating the likelihood Hessian we will use to sample from the approximated distribution around the MAP weights. At this time we do not include any NOTA data in the fitting procedure. For simple LA, we sample from the entire training dataset by the number datapoints in the dataset. For LAD, we fit to the full training dataset taking three times the samples than the size of the dataset. Finally, with the likelihood Hessian calculated (see equations 4.13 and 4.14 for how this is done), we perform a two-step logarithmic grid search using the validation set, to determine the best weight decay, γ^{-2} . We select the lowest weight decay that will achieve 96% of the MAP model's

Table 4.1. Model Accuracies

CIFAR-10				
Model	Det. Accuracy	MCD Stoch. Acc.	LAP Stoch Acc.	LAPMCD Stoch Acc
Undefended WRN-12-6	92.31%	92.27%	89.27%	89.35%
SRNA AdvMix WRN-12-6	92.2%	92.12	88.97%	88.99%
MRNA AdvMix WRN-12-6	91.04%	90.86%	87.62	88.22
CIFAR-100				
Undefended WRN-12-6	70.53%	70.73%	68.93%	67.87%
SRNA AdvMix WRN-12-6	69.86%	69.66%	67.22%	66.9%
MRNA AdvMix WRN-12-6	69.1%	69.62%	66.67%	66.75%

deterministic accuracy.

For both LA and MCD, we take 30 samples when making predictions. For LAD, we increase that number and take 50 samples (including when selecting γ^{-2}) to reduce variability in predictions. For attacks, as before, we set maximum iterations to 100 unless otherwise specified *higher*, such as in subordinate attacks within AutoAttack. All attacks are modified and executed according to one of two separate strategies as discussed in Section 4.1.5. The DWBT strategy for each attack is executed by saving the deterministic model before applying the LA, turning the dropout layers on for MCD, or both for LAD. We then create adversarial examples for each attack from this original deterministic model and make predictions using the stochastic version of the models. The EOT strategy required creating all new ‘generate_stochastic’ functions in each ART library attack [137]. In the EOT implementation of attacks, we use 30 samples per loss calculation in LA and MCD and 50 for LAD when calculating adversarial perturbation.

4.3.2 Results

Experimental results in tables 4.2 and 4.3 demonstrate a number of significant findings. We will first assess the robustness that results from converting an undefended (non-NOTA) model to the three versions of Bayesian Neural Networks, MCD, LA, and LAD.

Non-NOTA Stochastic Approaches

In general, the small amount of stochasticity that we introduce¹⁹ confers *some* improvement to robustness, particularly against the SquareAttack in CIFAR-10 where the LA alone brings ASR down from 67% to 31% in DWBT.²⁰ Additionally, stochasticity alone does appear to improve robustness against AutoPGD-DLR, particularly in the case of LAD. Stochasticity alone does not create a defense that is robust across the range of evaluated attacks.

Our experimental findings regarding non-NOTA stochastic models do not suggest that they are a credible defense when limiting noise to that which reduces accuracy a maximum of 4% of deterministic base model accuracy. However, our interest does not lie solely in stochasticity, but, instead in the combination of stochasticity with NOTA-defended base models. It is here that our findings get interesting. The difference in adversarial robustness observed between either deterministic NOTA or stochastic non-NOTA models and stochastic NOTA models is *stark* by contrast.

NOTA-Stochastic Combinations

In CIFAR-10, stochastic versions of both SRNA and MRNA AdvMixup result in dramatic reductions in ASR across all attacks. LAD SRNA gives the strongest results against the widely-held benchmark for robustness, untargeted AutoAttack, where it decreases ASR *below* the natural error rate of the model for both the L_2 bound of 0.5 and the L_∞ bound of 0.031.²¹ Both SRNA and MRNA performance against untargeted AutoAttack is commensurate with the top performance on Robust Bench for CIFAR-10 in both metrics [145]. Likewise, LAD SRNA is fully robust against AutoPGD-DLR, with its ASR reduced to less than half of the natural error rate for both EOT and DWBT. LAD SRNA posts the largest reduction of CWL₂ ASR to at or just above natural error rate. with a DWBT ASR of 11% in CIFAR-10.

Our CIFAR-100 results in table 4.3 show that the adversarial robustness observed for stochastic versions of AdvMixup in CIFAR-10 holds equally true for the more complicated

¹⁹We limited noise introduction to the highest amount that still maintained 96% of deterministic accuracy. This control is exerted through weight decay (γ^{-2}) selection during LA and LAD. MCD rarely resulted in any significant reduction in accuracy on its own and its noise, instead is set through the fixed dropout rate of 30%.

²⁰Any reporting and discussion of stochastic model performance will always take the worse ASR performance between the EOT or DWBT, as the attacker can always choose the most effective attack to use.

²¹These epsilons are specified as the benchmark distances over each L_p metric on Robust Bench [145], an up-to-date on-line listing of defense robustness.

Table 4.2. Stochastic CIFAR-10 Models vs. NOTA Stochastic Attacks
Bounds: Confidence:0, L_2 , $\epsilon = 0.5$, $L_\infty, \epsilon = 0.031$

Stochastic CIFAR-10 Models Vs NOTA Stochastic Attacks						
Models	C&W L_2	APGD DLR		Square Attk	AutoAttk, Untrgt	
<i>norm</i>	Conf:0	L_2	L_∞	L_∞	L_2	L_∞
DET No Def v reg	99%	87%	99%	67%	100%	100%
No Def MCD v DWBT	29%	84%	99%	46%	99%	100%
No Def MCD v EOT	100%	95%	98%	37%	99%	99%
No Def LAP v DWBT	26%	69%	96%	31%	94%	100%
No Def LAP v EOT	80%	97%	94%	24%	94%	97%
No Def LAD v DWBT	22%	71%	98%	30%	94%	100%
No Def LAD v EOT	93%	89%	95%	19%	84%	100%
DET SRNA AdvMix v NOTA Attks	9%	49%	88%	38%	67%	95%
SRNA MCD v NA-DWBT	7%	6%	8%	29%	7%	10%
SRNA MCD v NA-EOT	12%	3%	3%	21%	12%	29%
SRNA LAP v NA-DWBT	11%	6%	10%	32%	9%	14%
SRNA LAP v NA-EOT	16%	5%	2%	22%	17%	15%
SRNA LAD v NA-DWBT	11%	5%	5%	28%	8%	7%
SRNA LAD v NA-EOT	14%	4%	3%	14%	10%	11%
DET MRNA AdvMix v NOTA Attks	21%	49%	78%	61%	52%	89%
MRNA MCD v NA-DWBT	13%	9%	11%	41%	11%	17%
MRNA MCD v NA-EOT	20%	16%	25%	23%	22%	33%
MRNA LAP v NA-DWBT	13%	9%	16%	38%	15%	15%
MRNA LAP v NA-EOT	22%	6%	11%	27%	19%	18%
MRNA LAD v NA-DWBT	15%	9%	7%	23%	12%	10%
MRNA LAD v NA-EOT	15%	3%	0%	19%	13%	16%

dataset, CIFAR-100. This time, LAD versions of both SRNA and MRNA far surpass the reported robustness for the top CIFAR-100 defenses on Robust Bench against untargeted AutoAttack for both the L_2 and L_∞ norms. Additionally, both LAD SRNA and LAD MRNA reduce AutoPGD-DLR from L_2 ASRs of 81% (SRNA) and 92% (MRNA) and L_∞ ASRs of

Table 4.3. Stochastic CIFAR-100 Models vs. NOTA Stochastic Attacks
Bounds: Confidence:0, L_2 , $\epsilon = 0.5$, $L_\infty, \epsilon = 0.031$. DET stands for Deterministic Model

Stochastic CIFAR-100 Models Vs NOTA Stochastic Attacks						
Models	C&W L_2	APGD DLR		Square Attk	AutoAttk, Untrgtid	
<i>norm</i>	Conf:0	L_2	L_∞	L_∞	L_2	L_∞
DET No Def v Orig. Attk	99%	100%	100%	86%	100%	100%
No Def MCD v DWBT	37%	99%	97%	68%	97%	98%
No Def MCD v EOT	98%	99%	99%	65%	100%	100%
No Def LAP v DWBT	45%	96%	99%	73%	97%	98%
No Def LAP v EOT	82%	96%	99%	59%	97%	100%
No Def LAD v DWBT	46%	91%	96%	69%	98%	97%
No Def LAD v EOT	85%	80%	96%	52%	79%	97%
DET SRNA v NOTA Attk	43%	81%	86%	63%	88%	99%
SRNA MCD v NA-DWBT	30%	27%	32%	52%	27%	30%
SRNA MCD v NA-EOT	50%	53%	17%	37%	38%	34%
SRNA LAP v NA-DWBT	31%	24%	24%	46%	26%	30%
SRNA LAP v NA-EOT	45%	24%	15%	21%	30%	27%
SRNA LAD v NA-DWBT	25%	21%	25%	36%	22%	26%
SRNA LAD v NA-EOT	19%	9%	15%	9%	30%	26%
DET MRNA v NOTA Attk	48%	92%	96%	67%	92%	97%
MRNA MCD v NA-DWBT	36%	30%	56%	54%	30%	46%
MRNA MCD v NA-EOT	48%	48%	82%	45%	63%	78%
MRNA LAP v NA-DWBT	33%	27%	30%	56%	29%	37%
MRNA LAP v NA-EOT	35%	25%	46%	43%	32%	58%
MRNA LAD v NA-DWBT	36%	29%	29%	49%	29%	32%
MRNA LAD v NA-EOT	33%	16%	16%	21%	34%	30%

86% (SRNA) and 96% (MRNA) against the deterministic NOTA-defended models to *below natural error* for both DWBT and EOT in L_2 and L_∞ norms.

Attack Credibility and LAD Performance

It should also be noted that the DWBT and EOT versions of these attacks demonstrate highly effective successful attacks against the non-NOTA stochastic models in the table with upper 90s percentile ASRs. They are also moderately effective against MCD-MRNa and LA-MRNa. MCD-SRNA and LA-SRNA fare even better than MRNa. It is clear, however, that the combined distribution method, LAD, exhibits the maximum synergy between NOTa and the stochastic methods, with performance and robustness that far outstrip the previous state-of-the-art with respect to untargeted AutoAttack.

4.4 Discussion

Stochastic versions of Simple Reflexive NOTa-Aware Adversarial Mixup offer a promising way forward to adversarial robustness. However, due to the breadth of attacks we investigated, time constraints, and computing resources, there are logical avenues to continue to investigate to find weakness in these defenses. We performed these attacks with a maximum iteration of 100, this could be increased to 1000 or more, but at considerable computational cost especially for stochastic approaches. In some circumstances, such as AutoAttack and AutoPGD, however, default settings were used that had much higher maximum iterations as part of the algorithm, at times as high as 5000 iterations. Another avenue might be to consider new loss functions. One test that is worth investigating further would involve targeted attacks against each class—until failure of the defense is achieved for any example. Of note, higher so-called ‘confidences’ in Carlini and Wagner attacks have not resulted in higher attack success rates, however the profile of ASR vs confidence in those attacks should be added to this work. Additionally, showing a max ϵ distance vs ASR for several attacks would be useful to show that these defenses are not ‘tuned’ to the robust bench ϵ distances. We demonstrated performance across two L_2 ϵ s in Chapter 3 with DeepFool ϵ s of 0.1 and 0.5, showing robustness at each distance in the deterministic base models that were then used for this Chapter. Further work should demonstrate this robustness as well as the robustness to CW L_∞ and GSMA carries over and improves in derivative stochastic models such as MCD-SRNA, LA-SRNA, and LAD-SRNA.

Finally, in this Chapter, we began by investigating the effects of the injection of noise using Monte Carlo Dropout (MCD) and the Laplace Approximation (LA), into Deep Neural Networks (DNNs), limiting the noise to only enough to result in less than a 4% reduction

in accuracy. We also introduced a new stochastic noise sampling method, we call *Laplace Approximation with Dropout* (LAD), that samples from both the Bernouli and Gaussian distributions. Next, we combined these stochastic approaches with our NOTA defenses from Chapter 3 and found that the resulting Bayesian Neural Networks achieved adversarial robustness which exceeds reported state of the art robustness for CIFAR-10 and CIFAR-100. In the next Chapter, we will discuss how these findings along with the results of Chapters 2 and 3 combine to advance ML safety overall.

CHAPTER 5:

Discussion and Conclusion

We began in Chapter 1 by acknowledging a major issue of our age. AI, and ML specifically, is becoming ubiquitous, permeating every aspect of modern life. Increasingly, it has made its way into highly consequential systems that direct, protect and guide our lives, systems that can have life-altering or even life-ending repercussions when they malfunction. There are overt instances where this technology is chosen in these consequential contexts, but there are also obscure, seemingly inconsequential or tangential systems whose interactions can have unforeseen consequences. These forms of automation occur at speeds and scales which we can find difficult to comprehend or set proper oversight for, without careful prior systematic analysis and assessment. However, such assessment is an acknowledged shortcoming with industry leaders [25], even in systems that are consequential by design. Nevertheless, an assessment methodology for ML systems is a vital necessity to enable us to uncover the hidden hazards that exist, and mitigate or eliminate them before they result in catastrophe. In Section 1.2 we conducted a thought experiment about a notional DOD tank targeting system that uses an ML-based visual classification system, pointing out the many ways that such automation can have non-intuitive yet critical ramifications. Thus, perhaps even more critically, DOD has the same need to understand how an ML system alters and impacts the context into which it is introduced, and what can be done to ensure new hazards are prevented or mitigated.

However, most present efforts to incorporate safety into ML application development are focused on some combination of adhoc expert opinion, formal methods zeroed in on a particular aspect, or more narrowly on particular metrics. Absent in all of this is an overarching systemic assessment which situates and organizes how all of those facets should be brought together to create a unified case that a system is safe and trustworthy in its defined context. To respond to this need, in Chapter 2 we explore how we might identify and rectify hazards in the highly complex, consequential systems into which we introduce ML. Using system safety's System Theoretic Process Analysis (STPA), we conduct two case studies of consequential ML-enabled systems, discovering that the methodology, with some adaptations that we introduce, can be effective at ensuring system designers and stakeholders

foresee the foreseeable and implement effective controls to safeguard against losses. We find that STPA can provide a systematic, repeatable structured framework with which to reason about ML design decisions in context with the hazards and losses they may cause or prevent.

In support of making such structured safety cases, in Chapter 3 we focus in on a persistent problem in ML that threatens safety across industries and applications, the fragility of DNNs with respect to adversarial examples. In that chapter we introduced Adversarial Mixup, a naive NOTA augmentation defense that reformulates training on any dataset to include a none-of-the-above class, which it furnishes by creating PGD adversarial examples and ‘mixing’ them with the original image on the fly. Thus, the resulting NOTA class images *closely* surround and bound the benign class’ partition of input space and cause the DNN classifier to classify adversarial examples as the NOTA class. This initial defense was highly successful, however, after analysis of how the defense was working, we discovered that some standard modifications to attacks were necessary to adapt them so that they were not trivially defeated by the defense. After adapting the eight attacks, the naive AdvMixup did not fare as well in testing. We then analyzed our findings and developed two approaches to strengthening the defense, hoping to more completely seed the data-sparse regions between data-dense class partitions in input space. Our two adaptive forms of NOTA defense are Simple Reflexive NOTA Adversarial Mixup (SRNA) and Multi-modal Reflexive NOTA Adversarial Mixup (MRNA). Each of these approaches showed significant final improvement in adversarial robustness, but still did not exceed current state-of-the-art against benchmark attack tests, like AutoAttack.

Finally, noting the gap that remained after Chapter 3 in fully mitigating adversarial attacks, in Chapter 4 we demonstrate that extending deterministic NOTA models with BNN methods—like Monte Carlo Dropout, Laplace Approximation and a novel combination of the two, Laplace Approximation with Dropout (LAD), results in a surprising synergy that provides greater broad-spectrum adversarial attack resilience than either alone or what simple addition could account for. Particularly, the Laplace Approximation with Dropout SRNA (LAD-SRNA) performed best across the CIFAR-10 and CIFAR-100 datasets, exhibiting adversarial robustness exceeding the state of the art listed on Robust Bench [145]. Thus both Chapter 3 and 4 make contributions to mitigating a major subordinate concern within overall safety in machine learning—adversarial robustness. These techniques serve to provide a supporting assessment of safety from this particular hazardous state that would fit into the larger case

for safety that an STPA analysis could provide.

5.1 Implications

At a high level, this dissertation introduces a process that provides a standard logical framework to contextualize ML systems and identify and address potential system hazards. It offers a bridge to connect high-level goals like safety and the AI ethical principles to low level ML life-cycle design and implementation decisions. The resulting output of an STPA is a structured argument that addresses each potential hazardous state with an identified solution or control that can be separately assessed as to effectiveness. According to recent literature, these are current shortfalls within ML system development and would represent a significant contribution.

At the technical level, we accomplish a number of contributions. Overall, we reach and exceed state-of-the-art adversarial robustness against benchmark attacks for CIFAR-10 and CIFAR-100 datasets. Our best model, LAD-SRNA achieves attack success rates less than the natural error rate in both datasets. Ostensibly, this means we can achieve higher accuracy with higher robustness in classification systems. The applications abound, any DNN classifier which operates in a contested environment where deception is a concern can stand to benefit from employing our techniques. In DOD, target recognition systems, or imagery intelligence, sensors, all of these that are now employing DNN classifiers are potentially susceptible to attack and could benefit from the techniques we develop here.

Additionally, within the adversarial testing community, we adapt eight state-of-the-art attacks in two separate ways. First, we modify each with an option that enables it to be effective against NOTA defenses. Second, we adapt each to implement an EOT-based attack on stochastic models. This opens up these benchmarks for to provide robustness tests against future NOTA approaches, future stochastic approaches, and future combinations of the two. We also outline the general attack failure modes that enable a practitioner to adapt any other attack to be more effective against NOTA defenses.

5.2 Future Work

A number of studies should follow from this work. With respect to Chapter 2, there is a need to apply the STPA for social and ethical impact to a system that is less well established.

Such a study should ideally involve multiple groups kept separate. In one group there could be a ‘free form’ attempt to identify and mitigate risks specifically with no one knowing or using ideas from STPA, to establish a baseline from which to compare true STPA benefits. Another group trained on STPA and perhaps another using an alternative method, such as failure modes and effects analysis (FMEA). Additionally, the general applicability would be supported by another STPA for social and ethical impact on a large language model derivative system or a recommendation system such as the algorithm running YouTube. More work is necessary as well to more concretely explicate how, to what degree, and with what limitations values can be manifested and encoded through the STPA process.

Following Chapters 3 and 4, we believe a straight-forward extension is to acknowledge that the ideal condition in a NOTA-defended network is that all areas, unless trained with data otherwise, should be biased *toward the NOTA class*. There are a number of ways to accomplish this both in deterministic models and Bayesian models. As a direct result of this dissertation, it is apparent to the researchers that a NOTA augmentation process coupled with a biased NOTA network will result in the high accuracy, high adversarial robustness deterministic and stochastic DNNs that are needed to fully address this concern across large complex datasets. Finally, Chapters 3 and 4 should be followed up by extending these techniques to TinyImagenet and ImageNet, and establishing their state-of-the-art performance in those datasets as well.

List of References

- [1] J. Stavridis and J. Arquilla, “Weapons of mass disruption: A conversation on the future force, geopolitics and leadership,” 10 2021. The Naval Postgraduate School’s (NPS) Secretary of the Navy Guest Lecture (SGL) features retired U.S. Navy Adm. James G. Stavridis and NPS Distinguished Professor Emeritus John Arquilla discussing the topic in title [Online]. Available: <https://www.youtube.com/watch?v=p1XQfNv2PXU>
- [2] K. Barry. “Federal government opens safety defect investigation into Tesla autopilot crashes.” August 2021 [Online]. Available: <https://www.consumerreports.org/autonomous-driving/nhtsa-safety-defect-investigation-tesla-autopilot-crashes-a6996819019/>
- [3] B. d’Enquêtes et d’Analyses pour la sécurité de l’aviation civile, “Final report: On the accident on 1st June 2009 to the Airbus A330-203 registered F-GZCP operated by Air France flight AF 447 Rio de Janeiro - Paris,” Rep., July 2012 [Online]. Available: <https://www.bea.aero/docspa/2009/f-cp090601.en/pdf/f-cp090601.en.pdf>
- [4] “Summary of the FAA’s Review of the Boeing 737 MAX,” Federal Aviation Administration, online, Washington, DC, USA, November 2020 [Online]. Available: https://www.faa.gov/foia/electronic_reading_room/boeing_reading_room/media/737_RTS_Summary.pdf
- [5] G. Ceylan, I. A. Anderson, and W. Wood, “Sharing of misinformation is habitual, not just lazy or biased,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 4, p. e2216614120, 2023 [Online]. Available: <https://doi.org/10.1073/pnas.2216614120>
- [6] C. Garvie, “A Forensic Without the Science: Face Recognition in U.S. Criminal Investigations,” December 2022 [Online]. Available: <https://www.law.georgetown.edu/privacy-technology-center/publications/a-forensic-without-the-science-face-recognition-in-u-s-criminal-investigations>
- [7] M. Heikklä. “Dutch scandal serves as a warning for Europe over risks of using algorithms.” Mar. 2022 [Online]. Available: <https://tinyurl.com/dutchscandal>
- [8] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, “Dissecting racial bias in an algorithm used to manage the health of populations,” *Science*, vol. 366, no. 6464, pp. 447–453, 2019. Publisher: American Association for the Advancement

of Science _eprint: <https://science.sciencemag.org/content/366/6464/447.full.pdf> [Online]. Available: <https://doi.org/10.1126/science.aax2342>

- [9] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, “How we analyzed the compas recidivism algorithm,” *ProPublica*, May 2016 [Online]. Available: <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm>
- [10] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, “On the dangers of stochastic parrots: Can language models be too big?” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. New York, NY, USA: Association for Computing Machinery, 2021, p. 610–623 [Online]. Available: <https://doi.org/10.1145/3442188.3445922>
- [11] A. Kirilenko, A. S. Kyle, M. Samadi, and T. Tuzun, “The flash crash: High-frequency trading in an electronic market,” *The Journal of Finance*, vol. 72, no. 3, pp. 967–998, 2017 [Online]. Available: <http://www.jstor.org/stable/26652722>
- [12] *Autonomy in Weapon Systems*. Washington, DC, USA: Under Secretary of Defense (P), January 2023. Published: DOD Directive 3000.09 [Online]. Available: https://fas.org/irp/doddir/dod/d3000_09.pdf
- [13] “Our approach to responsible AI at Microsoft” [Online]. Available: <https://www.microsoft.com/en-us/ai/our-approach>
- [14] “Ethics guidelines for trustworthy AI | Shaping Europe’s digital future,” April 2019 [Online]. Available: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- [15] “Translation: Notice on Conducting the ‘Clear 2022 Comprehensive Governance of Algorithms’ Special Action” [Online]. Available: <https://digichina.stanford.edu/work/translation-notice-on-conducting-the-clear-2022-comprehensive-governance-of-algorithms-special-action/>
- [16] R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang, “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” *IBM Journal of Research and Development*, vol. 63, no. 4/5, pp. 4:1–4:15, 2019 [Online]. Available: <https://doi.org/10.1147/JRD.2019.2942287>
- [17] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, “Concrete problems in ai safety,” *arXiv preprint arXiv:1606.06565*, 2016.
- [18] R. I. J. Dobbe, “System Safety and Artificial Intelligence,” Feb. 2022. arXiv:2202.09292 [cs, eess] [Online]. Available: <http://arxiv.org/abs/2202.09292>

- [19] I. Raji and R. Dobbe, “Concrete problems in ai safety, revisited,” in *ICLR workshop on ML in the real world*, 2020.
- [20] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [21] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, “Are adversarial examples inevitable?” in *International Conference on Learning Representations*, 2019 [Online]. Available: <https://openreview.net/forum?id=r1IWUoA9FQ>
- [22] F. Tramer, N. Carlini, W. Brendel, and A. Madry, “On Adaptive Attacks to Adversarial Example Defenses,” in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds. Curran Associates, Inc., 2020, vol. 33, pp. 1633–1645 [Online]. Available: <https://proceedings.neurips.cc/paper/2020/file/11f38f8ecd71867b42433548d1078e38-Paper.pdf>
- [23] NIST, “Ai risk management framework concept paper,” December 2021 [Online]. Available: https://www.nist.gov/system/files/documents/2021/12/14/AI%20RMF%20Concept%20Paper_13Dec2021_posted.pdf
- [24] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, “Unsolved problems in ML safety,” *CoRR*, vol. abs/2109.13916, 2021 [Online]. Available: <https://arxiv.org/abs/2109.13916>
- [25] Rismani, Shalaleh, Shelby, Renee, Smart, Andrew, Jatho, Edgar, Kroll, Joshua, Moon, AJung, and Rostamzadeh, Negar, “From plane crashes to algorithmic harm: Applicability of safety engineering frameworks for responsible ml,” in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI ’23)*. New York, NY, USA: Association for Computing Machinery, 2023 [Online]. Available: <https://doi.org/10.1145/3544548.3581407>
- [26] N. S. C. on Artificial Intelligence, “Final report,” Mar. 2021 [Online]. Available: <https://www.nscai.gov/2021-final-report/>
- [27] A. Coston, A. Kawakami, H. Zhu, K. Holstein, and H. Heidari, “A Validity Perspective on Evaluating the Justified Use of Data-driven Decision-making Algorithms,” Feb. 2023. arXiv:2206.14983 [cs] [Online]. Available: <https://doi.org/10.48550/arXiv.2206.14983>
- [28] N. G. Leveson, “The use of safety cases in certification and regulation,” 2011 [Online]. Available: <https://dspace.mit.edu/handle/1721.1/102833>

- [29] A. Barton, E. Jatho, and V. Berzins, “Defending Against Adversarial Examples in Deep Neural Network Classifiers,” Naval Postgraduate School, Monterey, CA 93941, Rep. NPS-CS-21-002, 2021 [Online]. Available: <https://calhoun.nps.edu/handle/10945/68624>
- [30] S. Tangredi, “Sun Tzu versus AI: Why artificial intelligence can fail in great power conflict,” *Proceedings*, May 2021 [Online]. Available: <https://www.usni.org/magazines/proceedings/2021/may/sun-tzu-versus-ai-why-artificial-intelligence-can-fail-great-power>
- [31] N. Machiavelli and G. Bull, *The Prince*. Penguin Classics, London, England, 2003.
- [32] S. Tzu, *The Art of War*. Allandale Online Publishing, 2000 [Online]. Available: <https://archive.org/details/SunTzusArtOfWarJeanJosephMarieAmiot>
- [33] E. W. Jatho and J. A. Kroll, “AI: Too Fragile to Fight?” *Proceedings*, vol. 148, no. 2, Feb. 2022. Publisher: U.S. Naval Institute [Online]. Available: <https://www.usni.org/magazines/proceedings/2022/february/artificial-intelligence-too-fragile-fight>
- [34] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” *arXiv preprint arXiv:1312.6199*, 2013.
- [35] F. Tramer, N. Carlini, W. Brendel, and A. Madry, “On adaptive attacks to adversarial example defenses,” *arXiv preprint arXiv:2002.08347*, 2020.
- [36] C. Perrow, *Normal accidents: Living with high risk technologies*. New York: Basic Books, 1984.
- [37] “Collision between vehicle controlled by developmental automated driving system and pedestrian Tempe, Arizona March 18, 2018,” Highway Accident Report NTSB/HAR-19/03, National Transportation Safety Board, Washington, DC: USA", 2019 [Online]. Available: <https://www.nts.gov/investigations/accidentreports/reports/har1903.pdf>
- [38] M. L. Cummings, “Automation bias in intelligent time critical decision support systems,” in *Decision Making in Aviation*. Routledge, 2017, pp. 289–294.
- [39] C. C. Demchak and S. J. Tangredi, ““Sea Hacking” Sun Tzu,” in *AI at WAR: How Big Data, Artificial Intelligence and Machine Learning Are Changing Naval Warfare*, S. J. Tangredi and G. Galdorisi, Eds. Annapolis, MD, USA: Naval Institute Press, 2021, pp. 286–309.

- [40] J. T. Parken, Oliver. "M1 Abrams Tank Tested With Artificial Intelligence Targeting System." Feb. 2023 [Online]. Available: <https://www.thedrive.com/the-war-zone/m1-abrams-tank-tested-with-artificial-intelligence-targeting-system>
- [41] N. Strout. "Intelligence agency takes over Project Maven, the Pentagon's signature AI scheme." Apr. 2022. Section: Intel/GEOINT [Online]. Available: <https://www.c4isrnet.com/intel-geoint/2022/04/27/intelligence-agency-takes-over-project-maven-the-pentagons-signature-ai-scheme/>
- [42] C. Demarest. "ChatGPT can make short work of Pentagon tasks, Air Force CIO says." Mar. 2023. Section: AI [Online]. Available: <https://www.c4isrnet.com/artificial-intelligence/2023/03/01/chatgpt-can-make-short-work-of-pentagon-tasks-air-force-cio-says/>
- [43] M. C. Elish. "Sepsis Watch in Practice." October 2020 [Online]. Available: <https://points.datasociety.net/sepsis-watch-in-practice-5b06f88655fe>
- [44] V. Shankar, R. Roelofs, H. Mania, A. Fang, B. Recht, and L. Schmidt, "Evaluating machine accuracy on ImageNet," in *Proceedings of the 37th International Conference on Machine Learning* (Proceedings of Machine Learning Research), H. D. III and A. Singh, Eds. PMLR, 13–18 Jul 2020, vol. 119, pp. 8634–8644 [Online]. Available: <https://proceedings.mlr.press/v119/shankar20c.html>
- [45] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," 2020.
- [46] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" in *International Conference on Learning Representations*, 2019 [Online]. Available: <https://openreview.net/forum?id=r1lWUoA9FQ>
- [47] C. W. Fisher and B. R. Kingma, "Criticality of data quality as exemplified in two disasters," *Information & Management*, vol. 39, no. 2, pp. 109–116, 2001.
- [48] "Collision between U.S. Navy destroyer John S. McCain and tanker Alnic mc Singapore Strait, 5 miles northeast of Horsburgh lighthouse august 21, 2017," Marine Accident Report NTSB/MAR-19/01, National Transportation Safety Board, Washington, DC: USA", 2019 [Online]. Available: <https://www.nts.gov/investigations/accidentreports/reports/mar1901.pdf>
- [49] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983 [Online]. Available: [https://doi.org/https://doi.org/10.1016/0005-1098\(83\)90046-8](https://doi.org/https://doi.org/10.1016/0005-1098(83)90046-8)

- [50] P. Wärnestål, “Why human-centered design is critical to ai-driven services” [Online]. Available: <https://medium.com/swlh/why-human-centered-design-is-critical-to-ai-driven-services-e242a8067af>
- [51] N. Carlini, M. Jagielski, and I. Mironov, “Cryptanalytic extraction of neural network models,” 2020 [Online]. Available: <https://arxiv.org/abs/2003.04884>
- [52] C. E. Shannon, “Communication theory of secrecy systems,” *The Bell System Technical Journal*, vol. 28, no. 4, pp. 656–715, 1949 [Online]. Available: <https://doi.org/10.1002/j.1538-7305.1949.tb00928.x>
- [53] E. W. Jatho, L. O. Mailloux, E. D. Williams, P. McClure, and J. A. Kroll, “Concrete Safety for ML Problems: System Safety for ML Development and Assessment,” Feb. 2023. arXiv:2302.02972 [cs, eess] [Online]. Available: <https://doi.org/10.48550/arXiv.2302.02972>
- [54] G. E. Box, “Science and statistics,” *Journal of the American Statistical Association*, vol. 71, no. 356, pp. 791–799, 1976.
- [55] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, “Unsolved problems in ml safety,” *arXiv preprint arXiv:2109.13916*, 2021.
- [56] A. Tieleman and J. P. Meltzer. “The European Union AI Act: Next steps and issues for building international cooperation in AI.” June 2022 [Online]. Available: <https://tinyurl.com/EUAIAct>
- [57] “AI Risk Management Framework | NIST,” August 2022 [Online]. Available: https://www.nist.gov/system/files/documents/2022/08/18/AI_RMF_2nd_draft.pdf
- [58] I. D. Raji, E. M. Bender, A. Paullada, E. Denton, and A. Hanna, “AI and the everything in the whole wide world benchmark,” *arXiv preprint arXiv:2111.15366*, 2021.
- [59] M. M. Malik, “A Hierarchy of Limitations in Machine Learning,” 2020. eprint: 2002.05193 [Online]. Available: <https://arxiv.org/abs/2002.05193>
- [60] D. K. Mulligan, J. A. Kroll, N. Kohli, and R. Y. Wong, “This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology,” *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 119:1–119:36, Nov. 2019 [Online]. Available: <https://doi.org/10.1145/3359221>
- [61] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu, “Accountable Algorithms,” *University of Pennsylvania Law Review*, vol. 165, no. 3, 2017.

- [62] I. D. Raji, A. Smart, R. N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, and P. Barnes, “Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing,” *ACM Conference on Fairness, Accountability, and Transparency*, 2020.
- [63] A. F. Cooper, E. Moss, B. Laufer, and H. Nissenbaum, “Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning,” in *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, jun 2022 [Online]. Available: <https://doi.org/10.1145/3531146.3533150>
- [64] A. D. Selbst, D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi, “Fairness and abstraction in sociotechnical systems,” in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 59–68.
- [65] K. Holstein, J. Wortman Vaughan, H. Daumé, M. Dudik, and H. Wallach, “Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need?” in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI ’19)*. New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–16. event-place: Glasgow, Scotland Uk [Online]. Available: <https://doi.org/10.1145/3290605.3300830>
- [66] M. Madaio, L. Egede, H. Subramonyam, J. W. Vaughan, and H. Wallach, “Assessing the Fairness of AI Systems: AI Practitioners’ Processes, Challenges, and Needs for Support,” 2022. Issue: CSCW1 Pages: 1–26 Publication Title: Proceedings of the ACM on Human-Computer Interaction Volume: 6.
- [67] C. Wilson, A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli, “Building and Auditing Fair Algorithms: A Case Study in Candidate Screening,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. New York, NY, USA: Association for Computing Machinery, 2021, pp. 666–677. event-place: Virtual Event, Canada [Online]. Available: <https://doi.org/10.1145/3442188.3445928>
- [68] S.-M. Shin, S. H. Lee, S. K. I. Shin, I. Jang, and J. Park, “STPA-Based Hazard and Importance Analysis on NPP Safety I&C Systems Focusing on Human–System Interactions,” *Reliab. Eng. Syst. Saf.*, vol. 213, p. 107698, Sep. 2021.
- [69] N. G. Leveson, *Engineering a safer world: Systems thinking applied to safety*. The MIT Press, 2016.
- [70] D. Hendrycks, N. Carlini, J. Schulman, and J. Steinhardt, “Unsolved Problems in ML Safety,” June 2022. arXiv:2109.13916 [cs] [Online]. Available: <http://arxiv.org/abs/2109.13916>

- [71] N. Leveson, *STPA Handbook* [Online]. Available: https://psas.scripts.mit.edu/home/get_file.php?name=STPA_handbook.pdf
- [72] B. Antoine, “Systems Theoretic Hazard Analysis (STPA) applied to the risk review of complex systems : an example from the medical device industry,” Thesis, Massachusetts Institute of Technology, 2013. Accepted: 2013-07-09T19:30:13Z [Online]. Available: <https://dspace.mit.edu/handle/1721.1/79424>
- [73] J. M. Rising and N. G. Leveson, “Systems-Theoretic Process Analysis of space launch vehicles,” *Journal of Space Safety Engineering*, vol. 5, no. 3, pp. 153–183, Sep. 2018 [Online]. Available: <https://doi.org/10.1016/j.jsse.2018.06.004>
- [74] W. S. Lee, D. L. Grosh, F. A. Tillman, and C. H. Lie, “Fault Tree Analysis, Methods, and Applications A Review,” *IEEE Transactions on Reliability*, vol. R-34, no. 3, pp. 194–203, Aug. 1985. Conference Name: IEEE Transactions on Reliability [Online]. Available: <https://doi.org/10.1109/TR.1985.5222114>
- [75] D. H. Stamatis, *Failure mode and effect analysis: FMEA from theory to execution*. Quality Press, 2003.
- [76] N. Leveson, N. Dulac, K. Marais, and J. Carroll, “Moving Beyond Normal Accidents and High Reliability Organizations: A Systems Approach to Safety in Complex Systems,” *Organization Studies*, vol. 30, no. 2-3, pp. 227–249, Feb. 2009. Publisher: SAGE Publications Ltd [Online]. Available: <https://doi.org/10.1177/0170840608101478>
- [77] J.-L. Lions, L. Luebeck, J.-L. Fauquembergue, G. Kahn, W. Kubbat, S. Levedag, L. Mazzini, D. Merle, and C. O’Halloran, “Ariane 5 flight 501 failure report by the inquiry board,” 1996.
- [78] A. Albee, S. Battel, R. Brace, G. Burdick, J. Casani, J. Lavell, C. Leising, D. MacPherson, P. Burr, and D. Dipprey, “Report on the loss of the mars polar lander and deep space 2 missions,” 2000.
- [79] D. Martin, V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac, “Extending the Machine Learning Abstraction Boundary: A Complex Systems Approach to Incorporate Societal Context,” 2020. _eprint: 2006.09663.
- [80] J. Thomas, “Introduction to STPA,” 2021. Engineering Systems Lab, MIT [Online]. Available: <https://www.youtube.com/watch?v=2W-iqnPbhyc>
- [81] A. E. Kilby, “Algorithmic fairness in predicting opioid use disorder using machine learning,” in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT ’21)*. New York, NY, USA: Association for Computing

Machinery, 2021, p. 272 [Online]. Available: <https://doi.org/10.1145/3442188.3445891>

- [82] *Indiana Prescription Monitoring Program: Requestor User Support Manual*, Version 2.1 ed., Indiana Board of Pharmacy Prescription Monitoring Program (IN-SPECT). 9901 Linn Station Road Louisville, KY 40223, 2020, pp. 1–99 [Online]. Available: https://www.in.gov/pla/inspect/files/Narxcare_user_guide.pdf
- [83] *Texas Prescription Monitoring Program: Requestor User Support Manual*, Version 2.5 ed., Appriss Health. 9901 Linn Station Road Louisville, KY 40223, 2021, pp. 1–96.
- [84] Appriss. “N.C. Department of Health and Human Services Partners with Appriss Health to Provide Access to NarxCare Platform and Controlled Substances Reporting System Information Directly within Electronic Health Record and Pharmacy Management Systems.” Nov. 2018 [Online]. Available: <https://tinyurl.com/accesstoNarxCare2>
- [85] Bamboo Health. “NarxCare| Substance Abuse Software Solution.” 2022 [Online]. Available: <https://bamboohealth.com/solutions/narxcare/>
- [86] Appriss. “West Virginia Partners with Appriss Health to Enable Access to NarxCare Substance Use Disorder Platform Directly within Clinical Workflow for All Prescribers and Pharmacists State-Wide.” June 2018 [Online]. Available: <https://tinyurl.com/accesstoNarxCare>
- [87] A. E. Kilby, “The economics of pain management,” PhD Thesis, Department of Economics, Massachusetts Institute of Technology., Cambridge, MA, USA, 2016 [Online]. Available: <http://hdl.handle.net/1721.1/107321>
- [88] A. J. Fein, “The Top 15 U.S. Pharmacies of 2021: Market Shares and Revenues at the Biggest Companies,” 2021 [Online]. Available: <https://www.drugchannels.net/2022/03/the-top-15-us-pharmacies-of-2021-market.html>
- [89] R. Shelby, S. Rismeni, K. Hennee, A. Moon, N. Rostamzadeh, P. Nicholas, N. Yilla, J. Gallegos, A. Smart, E. Garcia, and G. Virk, “Sociotechnical harms: Scoping a taxonomy for harm reduction,” 2022. Submitted to the ACM Conference on Human Factors in Computing Systems (CHI) 2023, awaiting response.
- [90] A. Z. Jacobs, “Measurement and Fairness,” *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Mar. 2021. ISBN: 9781450383097 Publisher: ACM [Online]. Available: <https://doi.org/10.1145/3442188.3445901>

- [91] J. A. Kroll, "Accountability in Computer Systems," in *The Oxford Handbook of the Ethics of AI*, M. D. Dubber, F. Pasquale, and S. Das, Eds. Oxford: Oxford University Press, 2020. Section: 9 [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3608468
- [92] D. Martin, Jr, V. Prabhakaran, J. Kuhlberg, A. Smart, and W. S. Isaac, "Participatory Problem Formulation for Fairer Machine Learning Through Community Based System Dynamics," May 2020. _eprint: 2005.07572.
- [93] "Overdose Death Rates." Jan. 2022 [Online]. Available: <https://nida.nih.gov/research-topics/trends-statistics/overdose-death-rates>
- [94] J. D. Oliva, "Dosing Discrimination: Regulating PDMP Risk Scores," Rochester, NY, Jan. 2021 [Online]. Available: <https://doi.org/10.2139/ssrn.3768774>
- [95] E. P. Finley, A. Garcia, K. Rosen, D. McGeary, M. J. Pugh, and J. S. Potter, "Evaluating the impact of prescription drug monitoring program implementation: a scoping review," *BMC Health Services Research*, vol. 17, no. 1, p. 420, June 2017 [Online]. Available: <https://doi.org/10.1186/s12913-017-2354-5>
- [96] N. Dasgupta, K. Creppage, A. Austin, C. Ringwalt, C. Sanford, and S. K. Proescholdbell, "Observed transition from opioid analgesic deaths toward heroin," *Drug and Alcohol Dependence*, vol. 145, pp. 238–241, 2014 [Online]. Available: <https://doi.org/https://doi.org/10.1016/j.drugalcdep.2014.10.005>
- [97] "PDMP Resources," 2023 [Online]. Available: <https://www.commerce.alaska.gov/web/cbpl/ProfessionalLicensing/PrescriptionDrugMonitoringProgram/Resources.aspx>
- [98] P. D. M. P. Alaska, "Alaska PDMP Prescriber Report Card User Guide," 2023 [Online]. Available: commerce.alaska.gov/web/portals/5/pub/PDMPPrescriberReportCardUserGuide.pdf
- [99] T. Winters, "Guidance on PDMP Best Practices: Options for Unsolicited Reporting," US Office of Justice Programs, January 2014 [Online]. Available: <https://www.ojp.gov/pdffiles1/bja/247135.pdf>
- [100] K. K. Dineen and J. M. DuBois, "Between a rock and a hard place: Can physicians prescribe opioids to treat pain adequately while avoiding legal sanction?" *American Journal of Law and Medicine*, vol. 42, no. 1, p. 7–52, 2016 [Online]. Available: <https://doi.org/10.1177/0098858816644712>
- [101] D. D. Kim and N. Sibai, "The current state of opioid prescribing and drug enforcement agency (dea) action against physicians: an analysis of dea database 2004-2017," *Pain Physician*, vol. 23, no. 3, p. 297, 2020.

- [102] A. Kilby. “Algorithmic Fairness in Predicting Opioid Use Disorder using Machine Learning.” 2021 [Online]. Available: <https://ai.northeastern.edu/ai-events/expeditions-experiential-ai-anglela-kilby/>
- [103] U. S. G. A. GAO, Office, “Facial Recognition Technology: Federal Law Enforcement Agencies Should Better Assess Privacy and Other Risks | U.S. GAO,” June 2021 [Online]. Available: <https://www.gao.gov/products/gao-21-518>
- [104] K. J. Del Greco, “Facial Recognition Technology: Ensuring Transparency in Government Use — FBI,” June 2019 [Online]. Available: <https://www.fbi.gov/news/testimony/facial-recognition-technology-ensuring-transparency-in-government-use>
- [105] “Facial Identification Subcommittee.” January 2023. Last Modified: 2023-01-11T14:56-05:00 [Online]. Available: <https://www.nist.gov/organization-scientific-area-committees-forensic-science/facial-identification-subcommittee>
- [106] U. S. G. A. GAO, Office, “Facial Recognition Technology: Current and Planned Uses by Federal Agencies | U.S. GAO,” August 2021 [Online]. Available: <https://www.gao.gov/products/gao-21-526>
- [107] J. Brodtkin. “Clearview AI aims to put almost every human in facial recognition database.” Feb. 2022 [Online]. Available: <https://arstechnica.com/tech-policy/2022/02/clearview-ai-aims-to-put-almost-every-human-in-facial-recognition-database/>
- [108] “Clearview AI, inc. | Federal Award Recipient Profile | USAspending” [Online]. Available: <https://usaspending.gov/recipient/8cbe9b49-79af-44f4-3eb5-87296ee4a591-C/latest>
- [109] P. Grother, M. Ngan, K. Hanaoka, J. C. Yang, and A. Hom, “Ongoing Face Recognition Vendor Test (fvt) Part 1: Verification,” National Institute Of Standards And Technology (NIST), U.S. DEPARTMENT OF COMMERCE, Rep., December 2022 [Online]. Available: https://pages.nist.gov/frvt/reports/11/frvt_11_report.pdf
- [110] C. A. Goodhart and C. Goodhart, *Problems of monetary management: the UK experience*. Springer, 1984.
- [111] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language models are few-shot learners,” 2020 [Online]. Available: <https://arxiv.org/abs/2005.14165>

- [112] H. Valizadegan, M. J. S. Martinho, L. S. Wilkens, J. M. Jenkins, J. C. Smith, D. A. Caldwell, J. D. Twicken, P. C. L. Gerum, N. Walia, K. Hausknecht, N. Y. Lubin, S. T. Bryson, and N. C. Oza, “ExoMiner: A highly accurate and explainable deep learning classifier that validates 301 new exoplanets,” *The Astrophysical Journal*, vol. 926, no. 2, p. 120, feb 2022 [Online]. Available: <https://doi.org/10.3847/1538-4357/ac4399>
- [113] N. Savage, “How AI is improving cancer diagnostics,” *Nature*, vol. 579, pp. S14–S16, 03 2020 [Online]. Available: <https://doi.org/10.1038/d41586-020-00847-2>
- [114] E. Callaway, “‘it will change everything’: Deepmind’s ai makes gigantic leap in solving protein structures,” *Nature*, vol. 588, pp. 203–204, 12 2020 [Online]. Available: <https://doi.org/10.1038/d41586-020-03348-4>
- [115] A. Boles and P. Rad, “Voice biometrics: Deep learning-based voiceprint authentication system,” in *2017 12th System of Systems Engineering Conference (SoSE)*, 2017, pp. 1–6 [Online]. Available: <https://doi.org/10.1109/SYSOSE.2017.7994971>
- [116] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet Classification with Deep Convolutional Neural Networks,” in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
- [117] “Papers with Code - ImageNet Benchmark (Image Classification)” [Online]. Available: <https://paperswithcode.com/sota/image-classification-on-imagenet>
- [118] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 39–57.
- [119] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.
- [120] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, “Deepfool: a simple and accurate method to fool deep neural networks,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.
- [121] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, “Square attack: a query-efficient black-box adversarial attack via random search,” in *European Conference on Computer Vision*. Springer, 2020, pp. 484–501.
- [122] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” in *Security and Privacy (EuroS&P), 2016 IEEE European Symposium on*. IEEE, 2016, pp. 372–387.

- [123] F. Rosenblatt, “The perceptron: A probabilistic model for information storage and organization in the brain,” *Psychological Review*, vol. 65, pp. 386–408, 1958. Place: US Publisher: American Psychological Association [Online]. Available: <https://doi.org/10.1037/h0042519>
- [124] J. Krohn, G. Beyleveld, and A. Bassens, *Deep Learning Illustrated: A Visual, Interactive Guide to Artificial Intelligence*, 1st ed. Addison-Wesley Professional, 2019.
- [125] Y. LeCun, Y. Bengio, and others, “Convolutional networks for images, speech, and time series,” *The handbook of brain theory and neural networks*, vol. 3361, no. 10, p. 1995, 1995.
- [126] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based Learning Applied to Document Recognition,” in *Proceedings of the IEEE*. IEEE, 1998, vol. 86, pp. 2278–2324. Issue: 11.
- [127] A. Chakraborty, M. Alam, V. Dey, A. Chattopadhyay, and D. Mukhopadhyay, “Adversarial attacks and defences: A survey,” 2018.
- [128] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations (ICLR’15)*, 2015.
- [129] H. Zhang, Y. Yu, J. Jiao, E. Xing, L. El Ghaoui, and M. Jordan, “Theoretically principled trade-off between robustness and accuracy,” in *International Conference on Machine Learning*. PMLR, 2019, pp. 7472–7482.
- [130] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, “Measuring robustness to natural distribution shifts in image classification,” *arXiv preprint arXiv:2007.00644*, 2020.
- [131] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in *International Conference on Learning Representations (ICLR’15)*, 2015.
- [132] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, “Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition,” in *Proceedings of the 2016 acm sigsac conference on computer and communications security*, 2016, pp. 1528–1540.
- [133] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.

- [134] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards deep learning models resistant to adversarial attacks,” *arXiv preprint arXiv:1706.06083*, 2017.
- [135] F. Croce and M. Hein, “Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks,” in *International conference on machine learning*. PMLR, 2020, pp. 2206–2216.
- [136] F. Croce and M. Hein, “Minimally distorted adversarial examples with a fast adaptive boundary attack,” in *International Conference on Machine Learning*. PMLR, 2020, pp. 2196–2205.
- [137] M. Nicolae, M. Sinn, T. N. Minh, A. Rawat, M. Wistuba, V. Zantedeschi, I. M. Molloy, and B. Edwards, “Adversarial robustness toolbox v0.2.2,” *CoRR*, vol. abs/1807.01069, 2018 [Online]. Available: <http://arxiv.org/abs/1807.01069>
- [138] F. Croce, M. Andriushchenko, V. Schwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, “Robustbench: A Standardized Adversarial Robustness Benchmark,” *arXiv preprint arXiv:2010.09670*, 2020.
- [139] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” Nov. 2015. arXiv:1511.07528 [cs, stat] [Online]. Available: <https://doi.org/10.48550/arXiv.1511.07528>
- [140] A. Barton, “Defending Neural Networks Against Adversarial Examples,” PhD Thesis, 2018 [Online]. Available: <https://rc.library.uta.edu/uta-ir/handle/10106/27743>
- [141] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, “Adversarial attacks on deep-learning models in natural language processing: A survey,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 11, no. 3, pp. 1–41, 2020. Publisher: ACM New York, NY, USA.
- [142] S. Zagoruyko and N. Komodakis, “Wide residual networks,” *arXiv preprint arXiv:1605.07146*, 2016.
- [143] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” Jan. 2017. arXiv:1412.6980 [cs] [Online]. Available: <https://doi.org/10.48550/arXiv.1412.6980>
- [144] P. Foret, A. Kleiner, H. Mobahi, and B. Neyshabur, “Sharpness-Aware Minimization for Efficiently Improving Generalization,” Apr. 2021. arXiv:2010.01412 [cs, stat] [Online]. Available: <https://doi.org/10.48550/arXiv.2010.01412>

- [145] F. Croce, M. Andriushchenko, V. Schwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, “Robustbench: a standardized adversarial robustness benchmark,” *arXiv preprint arXiv:2010.09670*, 2020.
- [146] N. Carlini, A. Athalye, N. Papernot, W. Brendel, J. Rauber, D. Tsipras, I. Goodfellow, A. Madry, and A. Kurakin, “On Evaluating Adversarial Robustness,” Feb. 2019. arXiv:1902.06705 [cs, stat] [Online]. Available: <http://arxiv.org/abs/1902.06705>
- [147] N. Carlini and D. Wagner, “Towards evaluating the robustness of neural networks,” in *Security and Privacy (SP), 2017 IEEE Symposium on*. IEEE, 2017, pp. 39–57.
- [148] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal Adversarial Perturbations,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 86–94 [Online]. Available: <https://doi.org/10.1109/CVPR.2017.17>
- [149] A. Athalye, N. Carlini, and D. Wagner, “Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples,” *arXiv preprint arXiv:1802.00420*, 2018.
- [150] G. S. Dhillon and N. Carlini, “Erratum concerning the obfuscated gradients attack on stochastic activation pruning,” *CoRR*, vol. abs/2010.00071, 2020 [Online]. Available: <https://arxiv.org/abs/2010.00071>
- [151] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, “Towards robust neural networks via random self-ensemble,” 2018.
- [152] R. Pinot, L. Meunier, F. Yger, C. Gouy-Pailler, Y. Chevalere, and J. Atif, “On the robustness of randomized classifiers to adversarial examples,” *CoRR*, vol. abs/2102.10875, 2021 [Online]. Available: <https://arxiv.org/abs/2102.10875>
- [153] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif, “Theoretical evidence for adversarial robustness through randomization,” June 2019. arXiv:1902.01148 [cs, stat] [Online]. Available: <https://doi.org/10.48550/arXiv.1902.01148>
- [154] A. Géron, *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*. O’reilly, September 2019.
- [155] R. Bassett and J. Deride, “Maximum a posteriori estimators as a limit of Bayes estimators,” *Mathematical Programming*, vol. 174, no. 1-2, pp. 129–144, jan 2018 [Online]. Available: <https://doi.org/10.1007/s10107-018-1241-0>

- [156] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022 [Online]. Available: probml.ai
- [157] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*. Springer, 2006, vol. 4, no. 4.
- [158] H. Anwar, U. Qamar, and A. W. Muzaffar Qureshi, “Global optimization ensemble model for classification methods,” *The Scientific World Journal*, vol. 2014, 2014.
- [159] E. Daxberger, A. Kristiadi, A. Immer, R. Eschenhagen, M. Bauer, and P. Hennig, “Laplace redux - effortless Bayesian deep learning,” *CoRR*, vol. abs/2106.14806, 2021 [Online]. Available: <https://arxiv.org/abs/2106.14806>
- [160] J. Kirkpatrick, R. Pascanu, N. C. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell, “Overcoming catastrophic forgetting in neural networks,” *CoRR*, vol. abs/1612.00796, 2016 [Online]. Available: <http://arxiv.org/abs/1612.00796>
- [161] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, “Dropout: a simple way to prevent neural networks from overfitting,” *The journal of machine learning research*, vol. 15, no. 1, pp. 1929–1958, 2014.
- [162] Y. Gal and Z. Ghahramani, “Dropout as a bayesian approximation: Representing model uncertainty in deep learning,” in *international conference on machine learning*. PMLR, 2016, pp. 1050–1059.
- [163] G. S. Dhillon, K. Azizzadenesheli, J. D. Bernstein, J. Kossaifi, A. Khanna, Z. C. Lipton, and A. Anandkumar, “Stochastic activation pruning for robust adversarial defense,” in *International Conference on Learning Representations*, 2018 [Online]. Available: <https://openreview.net/forum?id=H1uR4GZRZ>
- [164] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” in *International Conference on Learning Representations*, 2018 [Online]. Available: <https://openreview.net/forum?id=Sk9yuql0Z>

Initial Distribution List

1. Defense Technical Information Center
Ft. Belvoir, Virginia
2. Dudley Knox Library
Naval Postgraduate School
Monterey, California



DUDLEY KNOX LIBRARY

NAVAL POSTGRADUATE SCHOOL

WWW.NPS.EDU

WHERE SCIENCE MEETS THE ART OF WARFARE