



## Artificial Intelligence for Defensive Cyber Operations

Major Jessica D.F. Bélanger

### JCSP 50

#### Master of Defence Studies

##### Disclaimer

Opinions expressed remain those of the author and do not represent Department of National Defence or Canadian Forces policy. This paper may not be used without written permission.

© His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2024.

### PCEMI n° 50

#### Maîtrise en études de la défense

##### Avertissement

Les opinions exprimées n'engagent que leurs auteurs et ne reflètent aucunement des politiques du Ministère de la Défense nationale ou des Forces canadiennes. Ce papier ne peut être reproduit sans autorisation écrite.

© Sa Majesté le Roi du chef du Canada, représenté par le ministre de la Défense nationale, 2024.

CANADIAN FORCES COLLEGE - COLLÈGE DES FORCES CANADIENNES

JCSP 50 - PCEMI n° 50  
2023 - 2024

Master of Defence Studies – Maîtrise en études de la défense

**Artificial Intelligence for Defensive Cyber Operations**

**Major Jessica D.F. Bélanger**

*“This paper was written by a candidate attending the Canadian Forces College in fulfilment of one of the requirements of the Course of Studies. The paper is a scholastic document, and thus contains facts and opinions which the author alone considered appropriate and correct for the subject. It does not necessarily reflect the policy or the opinion of any agency, including the Government of Canada and the Canadian Department of National Defence. This paper may not be released, quoted or copied, except with the express permission of the Canadian Department of National Defence.”*

*« La présente étude a été rédigée par un stagiaire du Collège des Forces canadiennes pour satisfaire à l'une des exigences du cours. L'étude est un document qui se rapporte au cours et contient donc des faits et des opinions que seul l'auteur considère appropriés et convenables au sujet. Elle ne reflète pas nécessairement la politique ou l'opinion d'un organisme quelconque, y compris le gouvernement du Canada et le ministère de la Défense nationale du Canada. Il est défendu de diffuser, de citer ou de reproduire cette étude sans la permission expresse du ministère de la Défense nationale. »*

## ACKNOWLEDGEMENTS

This work would not have been possible without the support of my academic advisor, Dr. J. Craig Stone, an Emeritus Associate Professor in the Department of Defence Studies at the Canadian Forces College, for allowing me to select this subject and his guidance throughout this study. I am deeply thankful to Dr. Marion McKeown, a professor at the Writing Centre of the Royal Military College of Canada Writing Centre, for her insightful feedback and all hours spent improving the quality of this study. Additionally, I am thankful to Dr. Matthew R. Kellett, the special advisor to the Director General of Cyber and Command and Control Information Systems Operations and Joint Force Cyber Component Commander, for providing valuable advice on employing artificial intelligence for cyber operations. Finally, thank you to my family members for their support and belief in me to achieve this journey.

## ABSTRACT

2024 is an excellent year for Artificial Intelligence (AI) in the Department of National Defence (DND) and Canadian Armed Forces (CAF). Becoming AI-enabled by 2030 is the commitment established by the newly published *DND and CAF AI Strategy*.<sup>1</sup> Subsequently, the new updated defence policy, *Our North, Strong and Free: A Renewed Vision for Canada's Defence*, announced the establishment of a CAF Cyber Command to improve the conduct of cyber operations.<sup>2</sup> As a result, this study answers the following question: How can the CAF Cyber Command leverage AI to conduct Defensive Cyber Operations (DCO) and become AI-enabled by 2030?

To answer this question, additional questions were raised up: Why should the CAF Cyber Command urgently become an AI-enabled organization by 2030? What is AI, and how does it work? How are AI models built? What are the risks of leveraging AI? How are AI models defended against malicious cyber actors? What regulations and compliance must be followed when leveraging AI? What are DCOs? What can AI do to improve the planning and conduct of DCOs?

To answer these questions, this study outlines eight recommendations to provide a comprehensive roadmap for the CAF Cyber Command to leverage AI to conduct Defensive Cyber Operations (DCOs) and become AI-enabled by 2030. Despite the current inherent challenges and obstacles of AI, these recommendations should initiate the discussion among the CAF Cyber Command leadership and staff insofar as the discussion begins now.

---

<sup>1</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy* (Ottawa: His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2024a), 1.

<sup>2</sup> National Defence, *Our North, Strong and Free: A Renewed Vision for Canada's Defence* (His Majesty the King in Right of Canada as represented by the Minister of National Defence, 2024c), 26.

## TABLE OF CONTENTS

Acknowledgements .....	i
Abstract .....	i
Table of Contents .....	ii
List of Figures .....	v
List of Tables .....	vi
List of abbreviations .....	vii
Introduction .....	1
<b>Question Statement and Thesis</b> .....	1
<b>Structure</b> .....	2
<b>Scope, Target Audience, and Terminology</b> .....	2
<b>Research Methodology</b> .....	3
Chapter 1 – Becoming AI-enabled by 2030 .....	5
<b>Section 1.1 – The Urgency of Becoming AI-Enabled by 2030</b> .....	5
AI Leveraged by CAF Adversaries .....	5
Cyber Defence Improved .....	5
<b>Section 1.2 – AI is not a Solution to All Problems</b> .....	6
Chapter 2 – AI Concepts and AI Model Components .....	8
<b>Section 2.1 – AI Model Layers and Components</b> .....	8
Layer 1: Data Layer .....	9
Layer 2: Machine Learning Framework and Algorithm Layer .....	12
Layer 3: Model / Intelligence Layer .....	12
Layer 4: Application Layer .....	17
AI Model Components Summary .....	17
<b>Section 2.2 – AI Concepts</b> .....	17
AI Subfields .....	17
AI Models Architectures .....	22
Machine Learning Algorithms .....	23
<b>Summary</b> .....	27
Chapter 3 – Leveraging AI .....	28

<b>Section 3.1 – Framework for Creating AI Models</b>	28
In-house Development versus Commercial-off-the-Shelf Product	28
<b>10-Step Framework</b>	29
Step 1 – Framing the Problem	31
Step 2 – Selecting and Preparing Datasets	33
Step 3 – Selecting the Model	36
Step 4 – Infrastructure Requirements	37
Step 5 – Training the Model	37
Step 6 – Validating the Model	37
Step 7 – Testing the Model	38
Step 8 – Deploying the Model	39
Step 9 – Monitoring and Maintaining the Model	39
Step 10 – Defending the Model	40
Summary	40
<b>Section 3.2 – Defending the Model</b>	40
Output Risks	40
Mitigation Measures Against Output Risks	41
Adversarial Machine Learning	43
Mitigation Measures Against AML	44
<b>Section 3.3 – Regulations and Compliance</b>	47
The DND and CAF AI Strategy Requirements	47
Legal, Inclusive, Ethical, and Safe	47
Interoperability and Partnerships	48
Directive on Automated Decision-Making	49
Guide on the use of Generative AI	50
Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems	51
Artificial Intelligence and Data Act	51
Pan-Canadian Artificial Intelligence Strategy	52
Summary	52
<b>Section 3.4 – XAI</b>	52

DARPA XAI .....	54
NIST 4 XAI Principles.....	54
XAI Techniques .....	55
Risks and Challenges Related to XIA.....	58
<b>Chapter 3 Summary</b> .....	59
Chapter 4 – AI for Defensive Cyber Operations.....	60
Section 4.1 – Defensive Cyber Operations .....	60
DCO Definitions .....	60
Key Requirements and Activities.....	61
Section 4.2 – Key Considerations When Leveraging AI for DCO .....	62
Current Trends of AI in Cybersecurity.....	62
Challenges.....	63
Section 4.3 – AI for Planning DCO .....	67
Cyber Threat Intelligence.....	68
Cyber Key Terrain.....	69
Prevention .....	71
Detection.....	72
Defeat.....	76
<b>Summary</b> .....	78
Chapter 5 - Conclusion and Recommendations .....	79
Section 5.1 – 8 Recommendations.....	79
Recommendation 1 – Must become AI-enabled Organization Soon.....	79
Recommendation 2 – A 10-Step Framework .....	80
Recommendation 3 – Reduce Output Risks .....	81
Recommendation 4 – Defending AI Models .....	81
Recommendation 5 – Be Aware of Regulations and Compliance .....	82
Recommendation 6 – Explore XAI Development .....	83
Recommendation 7 – Be Cognizant of Current Trends and Challenges of AI in DCOs .....	83
Recommendation 8 – Leveraging AI for DCO’s Activities and Managing Models .....	83
<b>Section 5.2 – Conclusion</b> .....	84
Bibliography .....	85

## LIST OF FIGURES

Figure 2.1 – AI Model for Cyber Security .....	9
Figure 2.2 – Model Updating Predictions for each Labeled Example in the Training Dataset .....	11
Figure 2.3 – Dataset Example: Labeled Example .....	11
Figure 2.4 – Dataset Example: Unlabeled Example .....	12
Figure 2.5 – Example of How Loss Functions Work .....	14
Figure 2.6 – Examples of High Loss (Left Model) and Low Loss (Right Model) .....	15
Figure 2.7 – Iterative Approach .....	15
Figure 2.8 Expert Systems Components .....	18
Figure 2.9 – Machine Learning and Common Algorithms .....	19
Figure 2.10 – ANN and DL Architecture Example .....	20
Figure 3. 1 – Chapter 1 Disadvantages of In-house ML Solutions and COTS .....	29
Figure 3.2 – Example of Dataset's Size .....	33
Figure 3. 3 – Slicing a Dataset into Three Sets for Training, Validation, and Testing .....	35
Figure 3.4 – Model for Cyber Security .....	36
Figure 3.5 – Validation and Testing Workflow .....	39
Figure 3.6 – Performance versus Explainability .....	53
Figure 3.7 – NIST XAI 4 Principles .....	55
Figure 3.8 – DARPA's 3 Strategies to Improve Explainability .....	56
Figure 3.9 – DARPA XAI Concept .....	57
Figure 4.1 – Application of AI in Cybersecurity .....	62
Figure 4.2 – Statistics of Attacks and Normal Behaviour in the CIC-IDS2017 Dataset .....	65
Figure 4.3 – Degree of Imbalance based on the Proportion of Minority Class .....	65
Figure 4.4 – Example of Extreme Imbalanced with a Distribution of 200 Negative Instances and 1 Positive Instance .....	66
Figure 4.5 – Steps of Downsampling and Upweighting .....	67
Figure 4.6 – The Cyber Domain Layers .....	70
Figure 4.7 – Pros and Cons of Supervised and Unsupervised ML for Cyber Threat Detection ...	74

## LIST OF TABLES

Table 2.1 – Data Terminology .....	9
Table 2.2 – Common Hyperparameters .....	16
Table 2.3 – Common Machine Learning Supervised Learning Algorithms .....	24
Table 2.4 – Common Deep Learning Algorithms .....	25
Table 3.1 – 10-Step Framework Summary .....	30
Table 3.2 – Datasets Splitting Categories and Definitions.....	35
Table 3.3 – Laws, Policies, Guidelines, Guidance, and Commitment for the Development and Implementation of AI-enabled Technologies .....	48
Table 4.1 – Defensive Cyber Operations' Activities for Planning and Execution of DCO .....	61
Table 4.2 – Possible AI Solutions to Improve Intrusion Detection System.....	73
Table 4.3 – Possible AI Solutions to Improve Malware Detection .....	75
Table 4.4 – Possible AI Solutions to Improve Incident Response .....	77
Table 5.1 – 10 Step Framework Summary .....	80
Table 5.2 – Laws, Policies, Guidelines, Guidance, and Commitment for the Development and Implementation of AI-enabled Technologies .....	82

## LIST OF ABBREVIATIONS

AI – Artificial Intelligence

AIA – Algorithm Impact Assessment

ANN – Artificial Neural Network

AML – Adversarial Machine Learning

APT – Advanced Persistent Threat

AWS – Amazon Web Services

BIC – Bio-inspired Computations

CNN – Convolutional Neural Network

COTS – Commercial-Off-The-Shelf

DCAIC – DND/CAF Artificial Intelligence Centre

DCOs – Defensive Cyber Operations

DCO-IDM – Defensive Cyber Operations – Internal Defensive Measures

DCO-RA – Defensive Cyber Operations – Response Action

DTO – Digital Transformation Officer

DL – Deep Learning

GANs – Generative Adversarial Networks

GBA+ – Gender-Based Analysis Plus

IDS – Intrusion Detection System

ISR – Intelligence Surveillance and Reconnaissance

GPT – Generative Pre-Trained Transformers

LLM – Large Language Model

LoE – Line of Effort

MalGAN – Malicious Generative Adversarial Networks

ML – Machine Learning

NLP – Natural Language Processing

RNN – Recurrent Neural Network

TTPs – Tactics, Techniques, and Procedures

VA – Vulnerability Assessment

VAE – Variational Autoencoders

XAI – Explainable Artificial Intelligence

## INTRODUCTION

2024 could be considered the birthday of Artificial Intelligence (AI) in the Department of National Defence (DND) and the Canadian Armed Forces (CAF). On 7<sup>th</sup> March 2024, the Deputy Minister, Bill Matthews, and the Chief of Defence Staff (CDS), General Wayne Eyre, announced the first-ever *DND and CAF Artificial Intelligence (AI) Strategy*.<sup>3</sup> About a month later, on 8<sup>th</sup> April 2024, Prime Minister Justin Trudeau and Defence Minister Bill Blair announced the defence policy update: *Our North, Strong and Free: A Renewed Vision for Canada's Defence*.<sup>4</sup> These two significant events will accelerate the DND and CAF's AI development and operationalization.

*The DND and CAF AI Strategy* engages the Defence Team to become AI-enabled by 2030 and presents a strategic alignment to achieve this vision through five lines of effort (LoE).<sup>5</sup> The updated defence policy recognizes that AI is shaping the character of conflict by revolutionizing the conduct of conventional armed conflicts and competition under the edge of war, especially in the cyber domain.<sup>6</sup> AI can significantly bolster the cyber capabilities of CAF adversaries by automating and amplifying their offensive cyber operations, disinformation and interference campaigns, and espionage activities.<sup>7</sup> On the other hand, the CAF can leverage AI to enhance its cyber capabilities and improve the conduct of deception and disinformation operations.<sup>8</sup> In addition to recognizing AI as a game changer, the updated defence policy announced the establishment of a CAF Cyber Command to improve the conduct of cyber operations.<sup>9</sup> The new CAF Cyber Command might become involved in the CAF's operationalization and defence of AI technologies since AI technologies are built and operated from the cyber domain. It is reasonable to assume that CAF adversaries will leverage AI to manoeuvre in cyberspace and potentially target the DND and CAF. Therefore, the CAF Cyber Command will also need to leverage AI technologies to counter its adversaries in cyberspace.

### Question Statement and Thesis

This study proposes the following question: How can the CAF Cyber Command leverage AI to conduct Defensive Cyber Operations (DCOs) and become AI-enabled by 2030? The CAF Cyber Command must leverage AI to improve the efficiency of key planning and conducting

---

<sup>3</sup> "DM/CDS Message: Launch of the DND/CAF Artificial Intelligence Strategy," last modified March 7, accessed April 15, 2024, <https://www.canada.ca/en/department-national-defence/maple-leaf/defence/2024/03/dm-cds-message-launch-dnd-caf-artificial-intelligence-strategy.html>.

<sup>4</sup> "Our North, Strong and Free: A Renewed Vision for Canada's Defence," last modified April 8, accessed April 15, 2024, <https://www.canada.ca/en/department-national-defence/news/2024/04/our-north-strong-and-free-a-renewed-vision-for-canadas-defence.html>.

<sup>5</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 1.

<sup>6</sup> National Defence, *Our North, Strong and Free: A Renewed Vision for Canada's Defence*, 9.

<sup>7</sup> *Ibid.*

<sup>8</sup> National Defence, *Our North, Strong and Free: A Renewed Vision for Canada's Defence* (His Majesty the King Right of Canada as represented by the Minister of National Defence, 2024c), 9.

<sup>9</sup> *Ibid.*, 26.

DCO activities by following a framework to create AI models while complying with the laws, regulations, and policies of the Government of Canada and DND/CAF and aligning with the vision of *The DND and CAF AI Strategy*. These recommendations will support the CAF Cyber Command's objective of becoming an AI-enabled organization by 2030, which is aligned with the ambitious commitment set by *The DND and CAF AI Strategy's* vision.

## Structure

This study is divided into five chapters. Each chapter answers specific questions that are related to the statement question. The first chapter answers the question: Why should the CAF Cyber Command urgently become an AI-enabled organization by 2030? It discusses the importance of becoming AI-enabled by 2030 and the impact of AI on DCOs. The second chapter answers the question: What is AI, and how does it work? It begins by defining AI. Then, it describes technical AI concepts and model layers and components. The third and fourth chapters are the core of this study. They are really about the “how” to leverage AI. Chapter 3 recommendations apply to any AI projects done by the DND and CAF. It answered the following questions: How are AI models built? What are the risks of leveraging AI? How can AI models be defended against malicious cyber actors? What regulations and compliance must be followed when leveraging AI? This chapter also proposes a framework for building AI models and identifies the CAF Cyber Command leadership’s responsibilities when leveraging AI. It explains the risks related to AI, describes the regulations and compliance that must be followed and discusses XAI, an emergent field that will help reduce risks and increase the transparency of AI models. Chapter 4 focuses on AI for DCOs and answers how AI can be leveraged to plan and conduct DCOs. Finally, chapter 5 summarizes all recommendations and concludes this study.

## Scope, Target Audience, and Terminology

This study aims to be understandable by any reader and provide reasonable explanations about the AI domain concepts. It acknowledges that AI is evolving quickly, and the proposed recommendations might rapidly become outdated. Furthermore, the recommendations do not cover all possible applications, techniques, and algorithms for DCOs, but only the most common at the time of writing.

The latest published Defence Policy, *Our North, Strong and Free: A Renewed Vision for Canada's Defence*, refers to establishing the CAF Cyber Command.<sup>10</sup> Therefore, the main target audience is the future CAF Cyber Command leadership and staff responsible for planning and conducting cyber operations in the CAF. The target audience also includes the leadership and staff of any CAF organization since this study focuses on AI solutions that could be applied to

---

<sup>10</sup> National Defence, *Our North, Strong and Free: A Renewed Vision for Canada's Defence* (His Majesty the King Right of Canada as represented by the Minister of National Defence, 2024c), ix.

any domain. Moreover, this study could be relevant to anyone interested in learning more about AI.

Here is important information on the terminology. First, the *Canadian Armed Forces (CAF) Joint Doctrine Note (JDN) 2017-02 Cyber Operations* categorized DCOs into two categories: DCO Internal Defensive Measures (DCO-IDM) and DCO Response Actions (DCO-RA)<sup>11</sup>. The scope of this study will only concern DCO-IDM, which will be referred to as DCO. Additionally, this study avoids mentioning “cyber security” when possible since, in the CAF doctrine, there is a distinction between “cyber security” and “cyber defence”:

To help understand the practical difference between cyber security and cyber defence, is to recognize that cyber defence requires a shift from network assurance (security) to mission assurance, where cyber defence is focused on network availability and data confidentiality and integrity. Cyber defence focuses on sensing, detecting, orienting, and engaging adversaries to outmanoeuvre them and assists commanders in completing their missions successfully.<sup>12</sup>

As a result, it will employ the specific vocabulary related to the CAF Cyber Operations doctrine, such as DCO, instead of cyber security. Nevertheless, most of the recommendations and concepts discussed are from the literature about cyber security and could, therefore, be applied to DCO.

## Research Methodology

The research methodology for this study were divided into four categories: DND and CAF publications, the CAF cyber operations doctrine, literature provided by key industry leaders in the AI domain such as IBM, Google, and Amazon Web Services (AWS), and peer-reviewed journal articles.

The starting point was DND and CAF's four key publications: *The DND and CAF Data Strategy*, *The CAF Digital Campaign Plan*, *The DND and CAF AI Strategy*, and the updated defence policy *Our North, Strong and Free: A Renewed Vision for Canada's Defence*.

The *Canadian Armed Forces (CAF) Joint Doctrine Note (JDN) 2017-02—Cyber Operations* was used to define DCOs. It was also used to identify key activities for planning and executing DCOs to find ways to leverage AI so the CAF can improve its conduct of DCOs. The downside of referring to the *JDN 2017-02 Cyber Operations* is that the doctrine was published seven years ago, and it might be outdated considering the evolution speed of the cyber domain. Nevertheless, this is the only CAF's official doctrine about cyber operations.

---

<sup>11</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations* (Her Majesty the Queen as represented by the Minister of National Defence, 2107), 4-6.

<sup>12</sup> *Ibid.*, 1-2.

The literature provided by key industry leaders, such as IBM, Google, AWS, and Gartner, was more appropriate for comprehensively explaining AI and machine learning technical concepts. For instance, the *Google Machine Learning Fundamental Course* provided abundant information about the functioning of AI models. This literature was mostly used to write Chapters 2 and 3.

Finally, many peer-reviewed journal articles were used to deepen concepts and propose key AI and cyber operations considerations. Since AI and cyber domains are evolving rapidly, the research tried to limit the use of articles published after 2020 and ideally after the end of 2022. That is because ChatGPT, a powerful natural language and content creator, became publicly available on November 30, 2022.<sup>13</sup> This advancement and access to the public to this technology was leveraged in many applications, such as the cyber domain.<sup>14</sup>

---

<sup>13</sup> Jim Euchner, "Generative AI," *Research-Technology Management* 66, no. 3 (-04-20, 2023), , 71. doi:10.1080/08956308.2023.2188861. <https://doi.org/10.1080/08956308.2023.2188861>.

<sup>14</sup> Utpal Chakraborty, Soumyadeep Roy and Sumit Kumar, *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World* (London: BPB Publications, 2023), 211.

## CHAPTER 1 – BECOMING AI-ENABLED BY 2030

### Section 1.1 – The Urgency of Becoming AI-Enabled by 2030

Several DND and CAF official publications, such as the *DND and CAF Data Strategy*, *The CAF Digital Campaign Plan*, *The DND and CAF AI Strategy*, and the latest defence policy updated *Our North, Strong and Free: A Renewed Vision for Canada's Defence*, acknowledge the urgency of becoming AI-enabled 2030. In addition to this acknowledge further arguments could be made to justify the integration of AI into cyber operations. The first argument is that AI technologies significantly impact the cyber domain threat landscape. The second argument is that the CAF Cyber Command will improve its posture by integrating AI technologies into its cyber defence.

#### AI Leveraged by CAF Adversaries

The improvement of AI technologies is changing the cyber threat landscape in two ways: increasing the number of cyber actors and improving the sophistication, speed, and scale of cyber attacks. First, AI is increasing the number of cyber actors by lowering the entry barriers for cyber attackers. The best example is that generative AI facilitates the writing of malicious codes.<sup>15</sup> For instance, by asking ChatGPT to create a code that encrypts folders, a malicious actor can use it to conduct a ransomware attack.<sup>16</sup> In addition to the increase in the number of cyber actors, their attack could increase in speed and scale when they leverage AI.<sup>17</sup> AI improves attack techniques; they are more sophisticated and newer.<sup>18</sup> Moreover, malicious actors use machine learning techniques to bypass traditional cybersecurity measures to reach their target.<sup>19</sup> Consequently, with AI's rapid evolution and cyber actors leveraging it, AI will change the character of conflict.<sup>20</sup> Nevertheless, AI can also mitigate traditional and AI-enabled cyber attacks.

#### Cyber Defence Improved

The CAF Cyber Command should integrate AI technologies to improve its cyber defence posture. Traditional cyber defence requires several systematic processes,<sup>21</sup> such as alert triage, vulnerability assessment, incident response, etc. Often, these tasks must be done manually and

<sup>15</sup> Christof Ebert and Maximilian Beck, "Artificial Intelligence for Cybersecurity," *IEEE Software* 40, no. 6 (Nov 2023). doi:10.1109/ms.2023.3305726. <https://ieeexplore-ieee-org.cfc.idm.oclc.org/document/10339105>.

<sup>16</sup> Chakraborty, *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World*, 225.

<sup>17</sup> Thanh Cong Truong, Quoc Bao Diep and Ivan Zelinka, "Artificial Intelligence in the Cyber Domain: Offense and Defense," *Symmetry* 12, no. 3 (Mar 01, 2020), 1. doi:10.3390/sym12030410. <https://doi.org/10.3390/sym12030410>.

<sup>18</sup> *Ibid.*

<sup>19</sup> Kamran Shaukat et al., "Performance Comparison and Current Challenges of using Machine Learning Techniques in Cybersecurity," *Energies* 13, no. 10 (May 15, 2020), 18. doi:10.3390/en13102509. <https://doi.org/10.3390/en13102509>.

<sup>20</sup> National Defence, *Our North, Strong and Free: A Renewed Vision for Canada's Defence*, 9.

<sup>21</sup> Ebert, "Artificial Intelligence for Cybersecurity," 30.

require several highly trained employees. AI technologies have the potential to reduce this burden.<sup>22</sup> AI can make cyber defence more flexible, adaptable, and robust.<sup>23</sup> Organizations that have not begun integrating AI to defend their systems are becoming more vulnerable, and the risk of being attacked is increasing. Traditional cybersecurity mostly relies on known attacks or adversaries' tactics, techniques, and procedures (TTPs) and therefore, it has difficulties in countering unknown attacks.<sup>24</sup> Consequently, cyber defenders must maximize AI to detect, prevent, and defeat various intrusions or attacks in real-time.<sup>25</sup> AI will facilitate automatization to detect and defeat an adversary inside a system.<sup>26</sup> Additionally, it can help find relationships between different vectors of attacks and improve predictive intelligence.<sup>27</sup> Moreover, AI can handle significant data volumes, such as triage of intrusion detection systems alerts.<sup>28</sup> Finally, automatic incident response, such as isolating a network segment or potentially compromised workstations, can be done with AI.<sup>29</sup> These are a few examples of how AI can improve the conduct of DCOs, which will be discussed in more detail in Chapter 4.

## Section 1.2 – AI Is Not a Solution to All Problems

Since AI is no longer science fiction, it fundamentally changes how cyber security and defence deal with malicious actors.<sup>30</sup> Nevertheless, until proven otherwise, AI will not change the fundamental idea behind each cyber attack where a “real world aggressor” aims at a “real world target.”<sup>31</sup> Therefore, the cyber domain will observe different combinations between AI-based attack, AI-based defence, non-AI-based attack, and non-AI-based defence.<sup>32</sup> An important point to remember is that in the foreseeable future, AI will not suddenly own all space in cyberspace but will share it with elements that are not AI.<sup>33</sup>

Furthermore, AI is not always the best solution for achieving objectives for cyber attackers and defence. This technology cannot fix all problems by itself. As mentioned in the *Pan-Domain Force Employment Concept (PFEC)*, leveraging AI requires a deliberate and

---

<sup>22</sup> *Ibid.*

<sup>23</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 1.

<sup>24</sup> *Ibid.*, 3.

<sup>25</sup> *Ibid.*

<sup>26</sup> Canadian Centre for Cyber Security, *Artificial Intelligence - ITSAP.00.040* (Communications Security Establishment, 2023).

<sup>27</sup> *Ibid.*

<sup>28</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 3.

<sup>29</sup> *Ibid.*

<sup>30</sup> Adewale Daniel Sontan and Segun Victor Samuel, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," *World Journal of Advanced Research and Reviews* 21, no. 2 (February 28, 2024), , 1725. doi:10.30574/wjarr.2024.21.2.0607. <https://doi.org/10.30574/wjarr.2024.21.2.0607>.

<sup>31</sup> Daniel Ventre, *Artificial Intelligence, Cybersecurity and Cyber Defence* (Hoboken: John Wiley & Sons, Incorporated, 2020), 168.

<sup>32</sup> *Ibid.*

<sup>33</sup> *Ibid.*, 166.

thoughtful approach to frame the problem properly.<sup>34</sup> AI is still emerging, and not all organizations can leverage this technology. Complex AI projects require significant human expertise, time, and financial resources.<sup>35</sup> Even if this study advocates the urgency of becoming AI-enabled, it acknowledges the challenges of leveraging AI and flaws in AI models. Chapter 4 will discuss the challenges of leveraging AI in more detail. To summarize, this chapter's objective was to establish the necessity and urgency of the CAF Cyber Command to become an AI-enabled organization by 2030. It also highlighted the importance of leveraging AI in the planning and conducting DCO. The next chapter will explain technical AI concepts.

---

<sup>34</sup> National Defence, *Pan-Domain Force Employment Concept - Prevailing in a Dangerous World* (His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2023), 26.

<sup>35</sup> "What is Artificial Intelligence (AI)?" last modified n.d., accessed April 24, 2024, <https://aws.amazon.com/what-is/artificial-intelligence/>.

## CHAPTER 2 – AI CONCEPTS AND AI MODEL COMPONENTS

Before exploring any AI concepts, it is necessary to understand what AI is. The literature on this field is abundant, as can be seen in the bibliography of this study. Notwithstanding the existence of several different definitions of AI, this study will refer to *The DND and CAF AI Strategy*'s definition: "For DND/CAF, AI is considered the capability of a computer to do things that are normally associated with human logic, such as reasoning, learning, and self-improvement."<sup>36</sup> This chapter will attempt to explain AI model layers and classify AI concepts; however, the classification of these concepts varies throughout the literature. Moreover, the novelty of this domain and its rapid evolution could render this proposed classification outdated. For this reason, this chapter intends to provide a general understanding of key concepts most applicable to DCOs. The first section will describe the key components of an AI model and explain its functions. Then, the second section of this chapter will present AI subfields, model architectures, and common algorithms.

### Section 2.1 – AI Model Layers and Components

An AI model is "a program trained on a set of data to recognize certain patterns or make certain decisions without further human intervention."<sup>37</sup> An AI model can make decisions or predictions autonomously by applying different algorithms to data inputs and then providing the expected output they are programmed to.<sup>38</sup> In other words, AI models aim to analyze relationships between complex data to extract knowledge for classification purposes or predict new data.<sup>39</sup> Some literature, such as Amazon Web Services, used representations of layers to describe the components of AI models and their architecture, which are (1) the data layer, (2) the feature layer, (3) the intelligence or model layer, and (4) the application layer.<sup>40</sup> Figure 2-1 provides an example of an AI model architecture's layers for cyber security applications. The layer's name and distribution differ slightly from AWS's AI application architecture. This model contains four layers. This section will describe each layer and its key components.

---

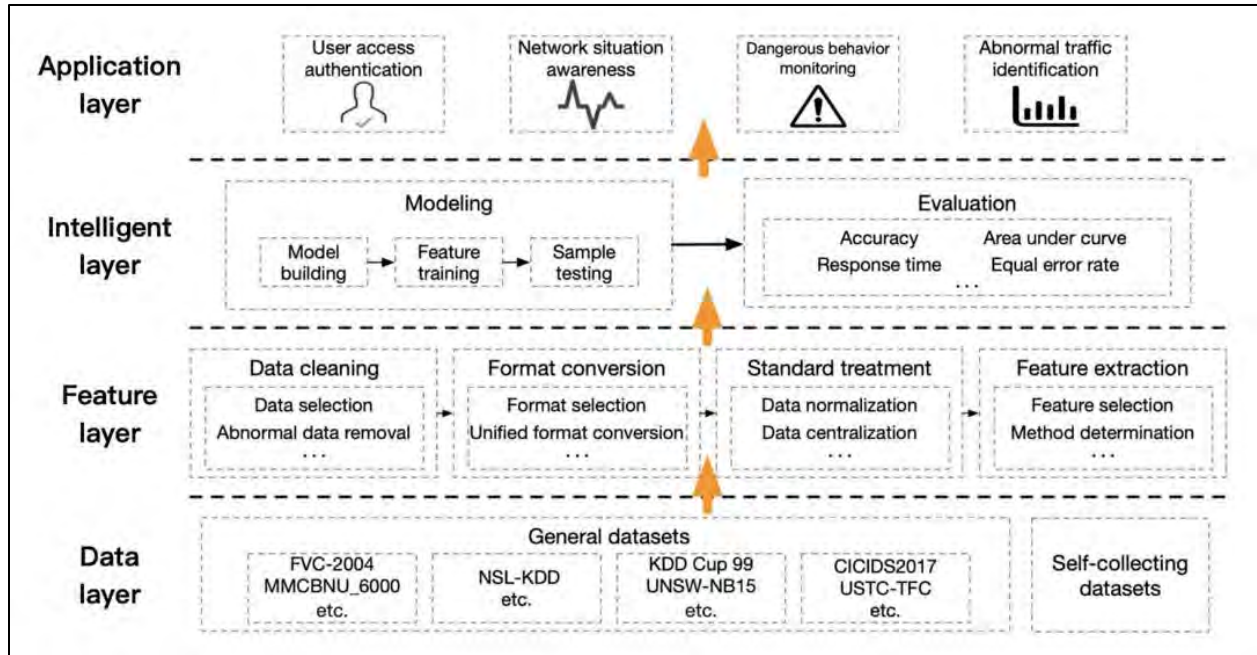
<sup>36</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 4.

<sup>37</sup> "What is an AI Model?", accessed Dec 20, 2023, <https://www.ibm.com/topics/ai-model>.

<sup>38</sup> *Ibid.*

<sup>39</sup> Mohiuddin Ahmed et al., "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence," *Studies in Computational Intelligence* 1025 (2022), 280. <https://doi.org/10.1007/978-3-030-96630-0>.

<sup>40</sup> "What is Artificial Intelligence (AI)?"



**Figure 2.1 – AI Model for Cyber Security**

Source: Zhimin Zhang et al., "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities," *Artificial Intelligence Review* 55, no. 2 (March 13, 2021), 1042. doi:10.1007/s10462-021-09976-0. <https://doi.org/10.1007/s10462-021-09976-0>.

#### Layer 1: Data Layer

Data could be considered the heart of each AI model, which forms the foundation layer of AI models.<sup>41</sup> Data are used for training, validating, and testing models. To achieve this objective, data must be collected and prepared for use by training algorithms, which will be detailed in the next section of this chapter. The AI field study has developed a specific data terminology. This study will use the *Google Machine Learning Foundational Course*<sup>42</sup> terminology. Table 2.1 defines key data terminology.

**Table 2.1 – Data Terminology**

Terminology	Definitions
<b>Datum / Data</b>	It can take different forms, such as text, numerical value, images, audio files, etc.
<b>Feature</b>	A feature is an input variable represented by (x). <sup>43</sup>
<b>Label</b>	The label is the model's predicted variable, usually represented by (y). <sup>44</sup>

<sup>41</sup> *Ibid.*

<sup>42</sup> "Machine Learning Crash Course - Validation Set: Another Partition," last modified November 2, accessed April 19, 2024, <https://developers.google.com/machine-learning/crash-course/validation/another-partition>.

<sup>43</sup> "Machine Learning Crash Course – Framing."

<sup>44</sup> *Ibid.*

<b>Example</b>	An example is an instance of data. <sup>45</sup>
<b>Labelled Example</b>	Labeled examples include both features and the label: {features, label}: (x, y). <sup>46</sup>
<b>Unlabelled Example</b>	Unlabeled examples contain only features: {features, ?}: (x, ?). <sup>47</sup>
<b>Dataset</b>	A dataset contains all labelled and unlabeled examples. The usual structure format is an Excel spreadsheet, a CSV file, or a JSON file. <sup>48</sup> The dataset will be split into three sets: training, validation, and testing.
<b>Training Set</b>	Sample of the original dataset used to train and fit the model.
<b>Validating Set</b>	After the model is trained, a validation set will be used for initial evaluation before using the testing set. The validation set can be used several iterations, which allows for tuning the model hyperparameters. Since the validation datasets are used several times, eventually, the evaluation becomes biased.
<b>Testing Set</b>	A sample of the original dataset is used against the model for an unbiased evaluation.

Sources: "Machine Learning Crash Course - Framing ," last modified July 18, accessed April 2, 2024, <https://developers.google.com/machine-learning/crash-course/framing/ml-terminology>. "What is the Difference between Test and Validation Datasets?" last modified August 14, accessed March 27, 2024, <https://machinelearningmastery.com/difference-test-validation-datasets/>. "Data Prep - Construct Your Dataset," last modified July 18, accessed April 3, 2024, <https://developers.google.com/machine-learning/data-prep/construct/construct-intro>. "Machine Learning Glossary - Training Set," last modified March 11, accessed April 3, 2024, [https://developers.google.com/machine-learning/glossary#training\\_set](https://developers.google.com/machine-learning/glossary#training_set).

Datasets contain individual examples made of features and labels for labelled examples or only features for unlabeled examples.<sup>49</sup> Each labelled or unlabeled example in a dataset will be fed to the model during its training, as per Figure 2.2 below.

---

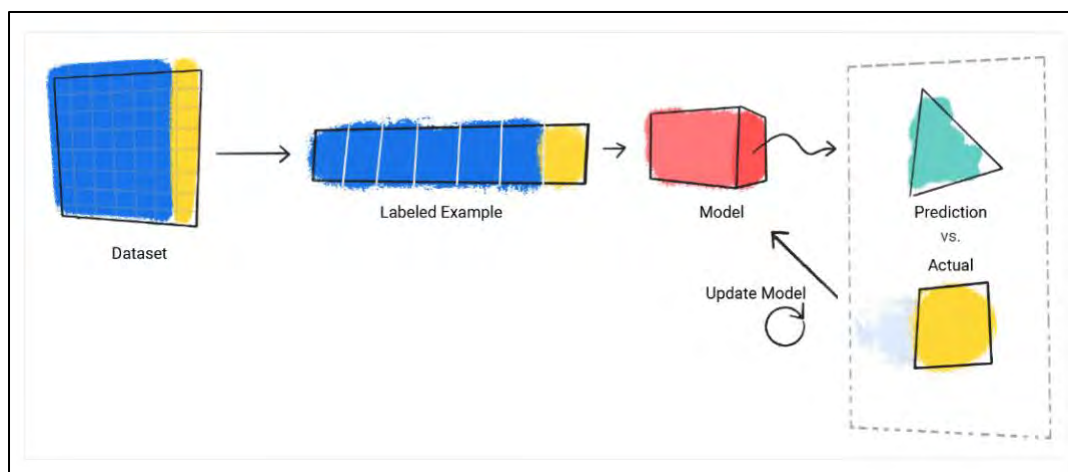
<sup>45</sup> *Ibid.*

<sup>46</sup> *Ibid.*

<sup>47</sup> *Ibid.*

<sup>48</sup> "Dataset," last modified n.d., accessed April 2, 2024, <https://www.databricks.com/glossary/what-is-dataset>.

<sup>49</sup> "What is Machine Learning?"



**Figure 2.2 – Model Updating Predictions for each Labeled Example in the Training Dataset**

Source: "What is Machine Learning?" last modified September 12, accessed April 2, 2024, <https://developers.google.com/machine-learning/intro-to-ml/what-is-ml>.

Each labelled example contains features and labels, whereas each unlabeled example contains only features. Meanwhile, the number of features can vary from one feature in a simple model to millions of features in more complex models.<sup>50</sup> Figures 2.3 and 2.4 below illustrate labelled and unlabeled examples that constitute datasets.

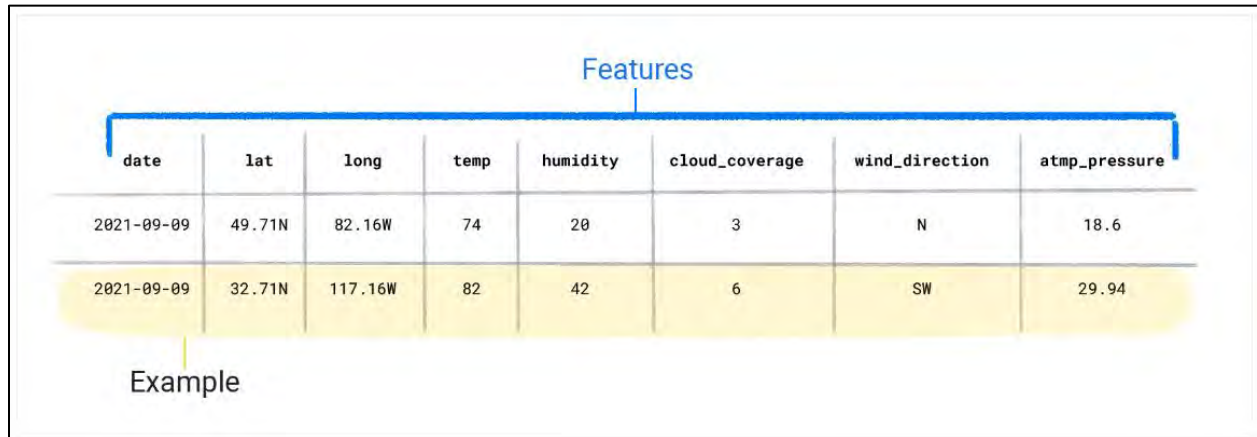
Features								Label
date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure	rainfall
2021-09-09	49.71N	82.16W	74	20	3	N	18.6	.01
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94	.23

Example

**Figure 2.3 – Dataset Example: Labeled Example**

Source: "What is Machine Learning?" last modified September 12, accessed April 2, 2024, <https://developers.google.com/machine-learning/intro-to-ml/what-is-ml>.

<sup>50</sup> "Machine Learning Crash Course – Framing."



The diagram shows a table with 8 columns: date, lat, long, temp, humidity, cloud\_coverage, wind\_direction, and atmp\_pressure. A blue bracket above the columns from temp to atmp\_pressure is labeled 'Features'. The second row of the table is highlighted in yellow and labeled 'Example' with a yellow arrow pointing to the 'date' column.

date	lat	long	temp	humidity	cloud_coverage	wind_direction	atmp_pressure
2021-09-09	49.71N	82.16W	74	20	3	N	18.6
2021-09-09	32.71N	117.16W	82	42	6	SW	29.94

**Figure 2.4 – Dataset Example: Unlabeled Example**

Source: "What is Machine Learning?" last modified September 12, accessed April 2, 2024, <https://developers.google.com/machine-learning/intro-to-ml/what-is-ml>.

During supervised training, which will be discussed in more detail in the next sections, the model is trained on labelled examples and then tested to predict the label on unlabeled examples.<sup>51</sup>

#### Layer 2: Machine Learning Framework and Algorithm Layer

This layer contains the machine learning framework for building and training the AI model. Common machine learning frameworks are TensorFlow, PyTorch, and scikit-learn.<sup>52</sup>

#### Layer 3: Model / Intelligence Layer

This layer contains the model structure where the AI model is implemented and trained with training algorithms and datasets.<sup>53</sup> The common components of this layer are the parameters, loss function, optimizer, and hyperparameters, which will be detailed in the following lines.

#### *Parameters*

Parameters are internal variables learned by the model during its training. Common parameters are the weight and biases for artificial neural networks and deep learning.<sup>54</sup> Parameters are adjusted during the training and validation steps to improve the model's performance and reduce the loss function.<sup>55</sup>

<sup>51</sup> "Machine Learning Crash Course – Framing."

<sup>52</sup> "What is Artificial Intelligence (AI)?"

<sup>53</sup> *Ibid.*

<sup>54</sup> "Machine Learning Glossary - Parameter," last modified March 11, accessed April 24, 2024, <https://developers.google.com/machine-learning/glossary#parameter>.

<sup>55</sup> "What is Artificial Intelligence (AI)?"

### *Loss Function*

The loss function measures the difference between the model's actual and expected predictions.<sup>56</sup> This component is responsible for measuring the performance during the training and provides feedback to the training algorithm.<sup>57</sup> The cost function represents the average of all loss function values.<sup>58</sup> When building an AI model, the aim is to minimize the loss function values, which means the model performs well.<sup>59</sup> Figure 2.5 provides an example of where the loss function works in an AI model and how it interacts with other components by updating the parameters. The model aims to minimize the loss, which is referred to as empirical risk minimization.<sup>60</sup>

---

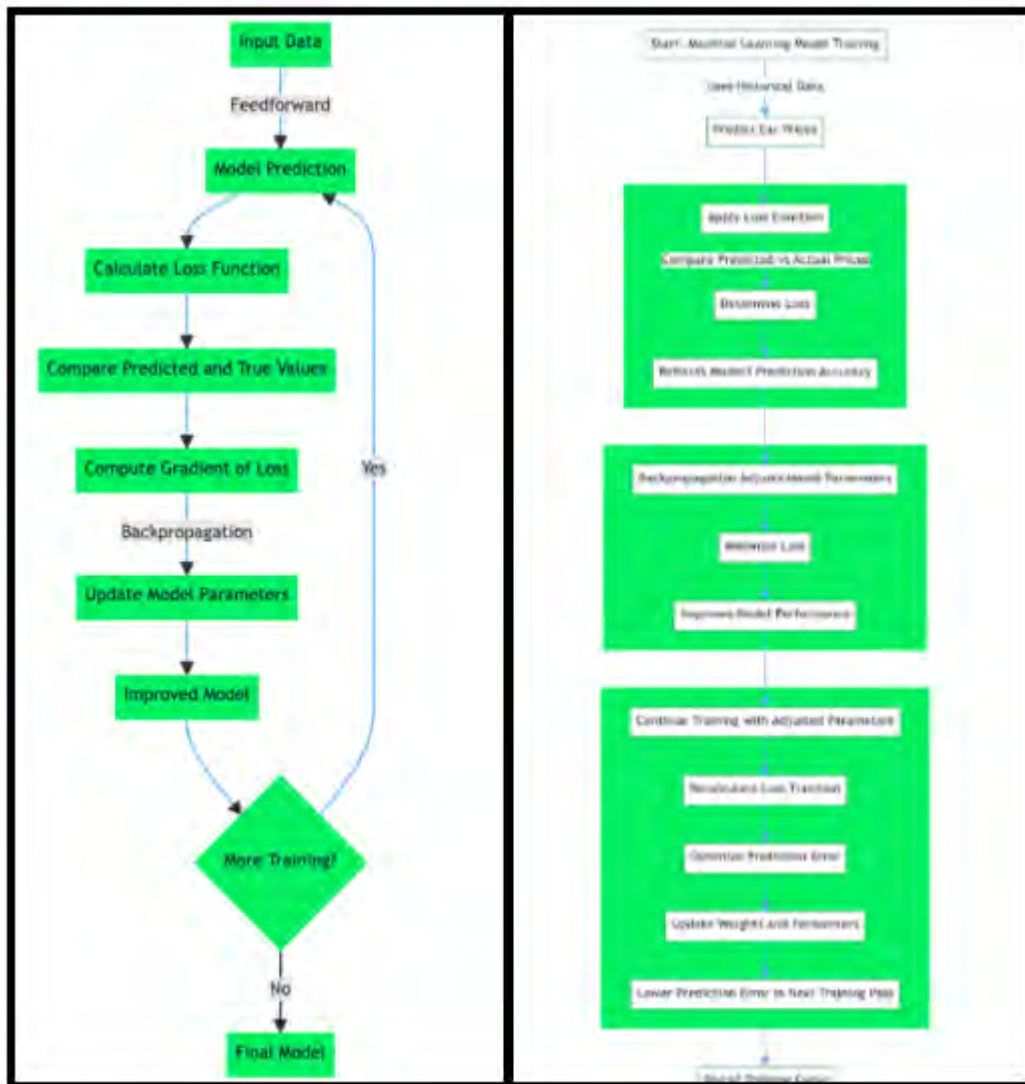
<sup>56</sup> "The 7 most Common Machine Learning Loss Functions ," last modified June 30, accessed March 7, 2024, <https://builtin.com/machine-learning/common-loss-functions>.

<sup>57</sup> "Machine Learning Crash Course - Descending into ML," last modified July 18, accessed April 2, 2024, <https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss>.

<sup>58</sup> "The 7 most Common Machine Learning Loss Functions."

<sup>59</sup> *Ibid.*

<sup>60</sup> "Machine Learning Crash Course - Descending into ML."

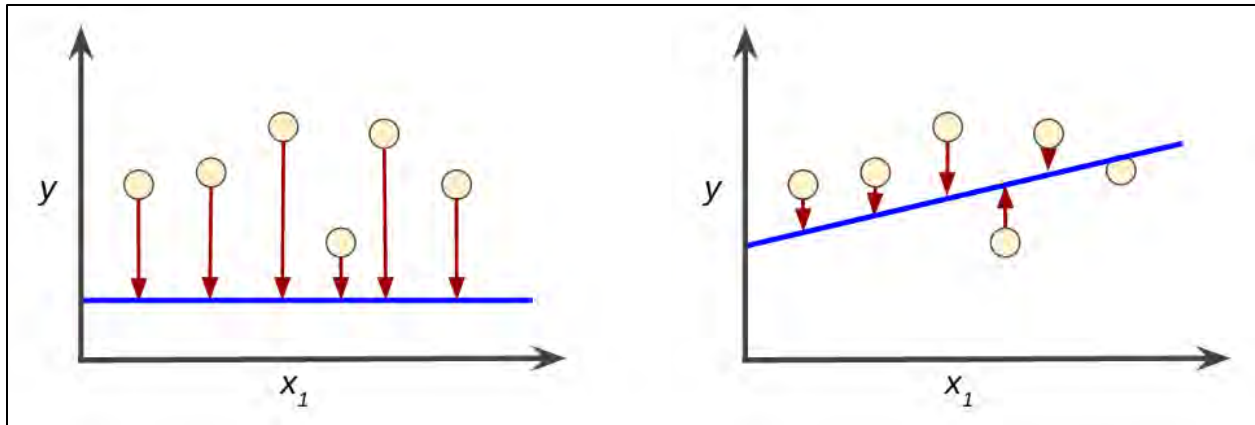


**Figure 2.5 – Example of How Loss Functions Work**

Source: "Loss Functions in Machine Learning Explained," accessed March 7, 2024, <https://www.datacamp.com/tutorial/loss-function-in-machine-learning>.

Figure 2.6 provides an example of loss reduction between the two models. The arrows represent the loss, and the blue lines represent the predictions.<sup>61</sup> The left model has a higher loss than the right model. The objective is to obtain a smaller loss function as possible. The model's parameters will be adjusted to reduce the loss function.

<sup>61</sup> "Machine Learning Crash Course - Descending into ML."

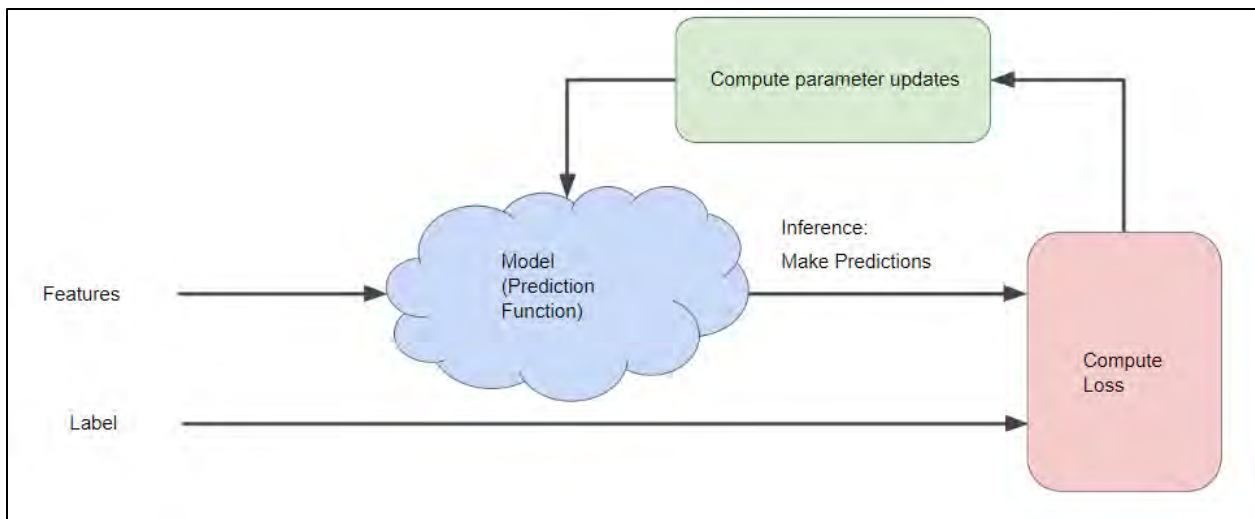


**Figure 2.6 – Examples of High Loss (Left Model) and Low Loss (Right Model)**

Source: "Machine Learning Crash Course - Reducing Loss," last modified November 16, accessed April 23, 2024, <https://developers.google.com/machine-learning/crash-course/reducing-loss/an-iterative-approach>.

### Optimizers

Optimizers are methods and algorithms that adjust parameters to reduce the loss function.<sup>62</sup> The iterative approach is an example of an optimizer method which consistently adjusts the model's parameters using an iterative trial-and-error process, as shown in Figures 2.7.<sup>63</sup>



**Figure 2.7 – Iterative Approach**

Source: "Machine Learning Crash Course - Reducing Loss."

<sup>62</sup> "What is Artificial Intelligence (AI)?"

<sup>63</sup> "Machine Learning Crash Course - Reducing Loss."

Common algorithms that optimize and adjust parameters include gradient descent and the Adaptive Gradient Algorithm (AdaGrad).<sup>64</sup>

### *Hyperparameters*

Finally, hyperparameters are configuration variables that modify the model structure, functions, and performance.<sup>65</sup> A way to differentiate hyperparameters from parameters is that the first are external configurations, and the latest are internal configurations automatically obtained during training.<sup>66</sup> Common hyperparameters are defined in Table 2.2 below.

**Table 2.2 – Common Hyperparameters**

<b>Hyperparameters</b>	<b>Definition</b>
<b>Learning Rate</b>	This is a key hyperparameter. It is used to adjust the gradient descent algorithm.
<b>Learning Rate Decay</b>	This hyperparameter reduces the learning rate gradually over time to accelerate the learning process.
<b>Momentum</b>	This hyperparameter will improve and accelerate the learning process by indicating the right direction or the next step to the algorithm.
<b>Neural Network Nodes</b>	This is the number of nodes in each hidden layer.
<b>Neural Network Layers</b>	This is the number of hidden layers in a neural network.
<b>Mini-batch Size</b>	It is a set of training examples to compute the gradient during one iteration.
<b>Epochs</b>	This is the number of times the entire training set is shown to the model during its training.
<b>Eta</b>	This hyperparameter is used to prevent overfitting.

Sources: "What is Hyperparameters Tuning?"

"Machine Learning Crash Course - Reducing Loss"

Modifying hyperparameters will improve the model's accuracy and performance. This process is usually called hyperparameter tuning, and it can be done automatically or manually. Algorithms can be used for hyperparameter tuning, and the common techniques are Bayesian optimization, grid search, and random search.<sup>67</sup>

<sup>64</sup> "What is Artificial Intelligence (AI)?"

<sup>65</sup> "What is Hyperparameters Tuning?" last modified n.d., accessed March 26, 2024, <https://aws.amazon.com/what-is/hyperparameter-tuning/>.

<sup>66</sup> *Ibid.*

<sup>67</sup> "What is Hyperparameters Tuning?"

## Layer 4: Application Layer

This layer is where the user interacts with the AI model.<sup>68</sup> It contains the input data and output predictions. Usually, the user provides input that could take many forms, such as text, videos, images, or a question in a prompt. Afterwards, the user expects an output from the AI model. The *raison d'être* for each AI model is their output prediction.

## AI Model Components Summary

In summary, the overall AI model architecture can be divided into four layers: data layer, machine learning framework and algorithms layer, model/intelligence layer, and application layer. This division of AI model architecture helps understand how the AI model functions. This knowledge will be useful in Chapter 3 to explain the proposed 10-step framework for building AI models.

## Section 2.2 – AI Concepts

### AI Subfields

The most common AI subfields, also called branches, are machine learning, deep learning, natural language processing, expert systems, robotics, machine vision, and speech recognition.<sup>69</sup> This chapter will describe the most relevant subfields for DCO: expert systems, machine learning, artificial neural networks, deep learning, and natural language processing.

### *Expert Systems*

Expert systems are often used to assist with decision-making and complex problems.<sup>70</sup> They have three components: a knowledge base, an inference engine, and a user interface, as shown in Figure 2-9. The knowledge base contains all information about a specific field and if-then rules.<sup>71</sup> The inference engine solves problems and makes decisions.<sup>72</sup> The last part, the user interface, allows interactions with the system.<sup>73</sup>

---

<sup>68</sup> "What is Artificial Intelligence (AI)?"

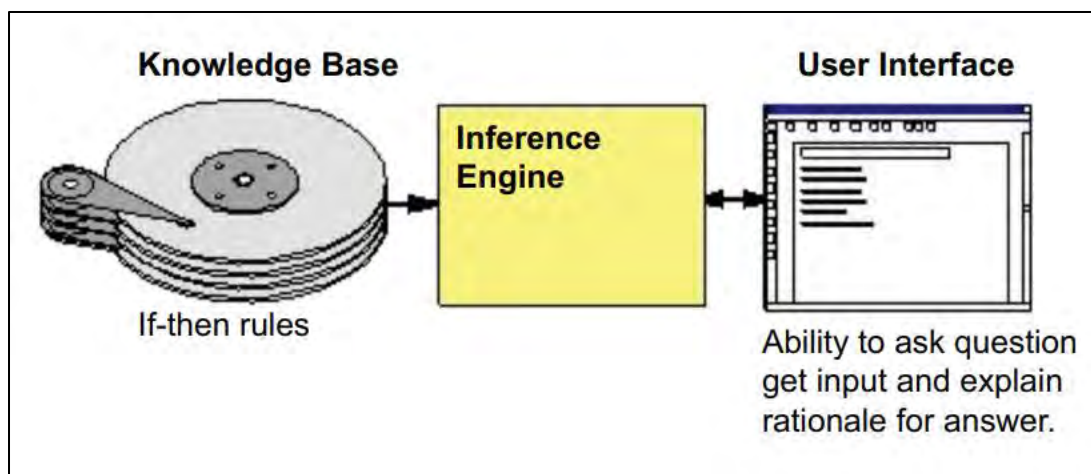
<sup>69</sup> Tolga Akcay, *The AI Compass - Welcome to the World of Artificial Intelligence* (United States: , 2021), 22.

<sup>70</sup> Rammanohar Das and Raghav Sandhane, "Artificial Intelligence in Cyber Security," *Journal of Physics: Conference Series* 1964, no. 4 (Jul 23, 2021), 6. doi:10.1088/1742-6596/1964/4/042072. <https://doi.org/10.1088/1742-6596/1964/4/042072>.

<sup>71</sup> Asefeh Asemi, Andrea Ko and Mohsen Nowkarizi, "Intelligent Libraries: A Review on Expert Systems, Artificial Intelligence, and Robot," *Library Hi Tech* 39, no. 2 (June 6, 2020), 415. doi:10.1108/lht-02-2020-0038. <https://doi.org/10.1108/LHT-02-2020-0038>.

<sup>72</sup> Daniel Ventre, *Artificial Intelligence, Cybersecurity and Cyber Defence* (Hoboken: John Wiley & Sons, Incorporated, 2020), 56.

<sup>73</sup> *Ibid.*



**Figure 2.8 Expert Systems Components**

Source: Asefeh Asemi, Andrea Ko and Mohsen Nowkarizi, "Intelligent Libraries: A Review on Expert Systems, Artificial Intelligence, and Robot," *Library Hi Tech* 39, no. 2 (June 6, 2020), , 416. doi:10.1108/lht-02-2020-0038. <https://doi.org/10.1108/LHT-02-2020-0038>.

Expert Systems are limited to their area of expertise and are not capable of considering the context in their analysis.<sup>74</sup> Therefore, other AI subfields are being developed to counter this issue.

#### *Machine Learning (ML)*

ML is a game changer in AI since it “[...] gives computers the ability to learn and improve from experiences without being explicitly programmed.”<sup>75</sup> Unlike conventional software or expert systems, which must be programmed with predefined rules to obtain an output, ML uses algorithms that can learn without human intervention.<sup>76</sup> Their goal is to learn automatically to make decisions.<sup>77</sup> ML relies on mathematical techniques that can extract information from the data, identify patterns and propose conclusions.<sup>78</sup> The ML algorithms will analyze datasets from which an ML model will be developed, and this model will incorporate all the knowledge learned.<sup>79</sup> Then, the model will use this knowledge learned to provide the right output associated with the input, which is usually unseen data.<sup>80</sup> ML uses three learning methods:

<sup>74</sup> *Ibid.*

<sup>75</sup> Ahmed, "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence," 43.

<sup>76</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 4.

<sup>77</sup> Giovanni Apruzzese et al., "The Role of Machine Learning in Cybersecurity," *Digital Threats: Research and Practice* 4, no. 1 (March 7, 2023), 4. doi:10.1145/3545574. <http://arxiv.org/abs/2206.09707>.

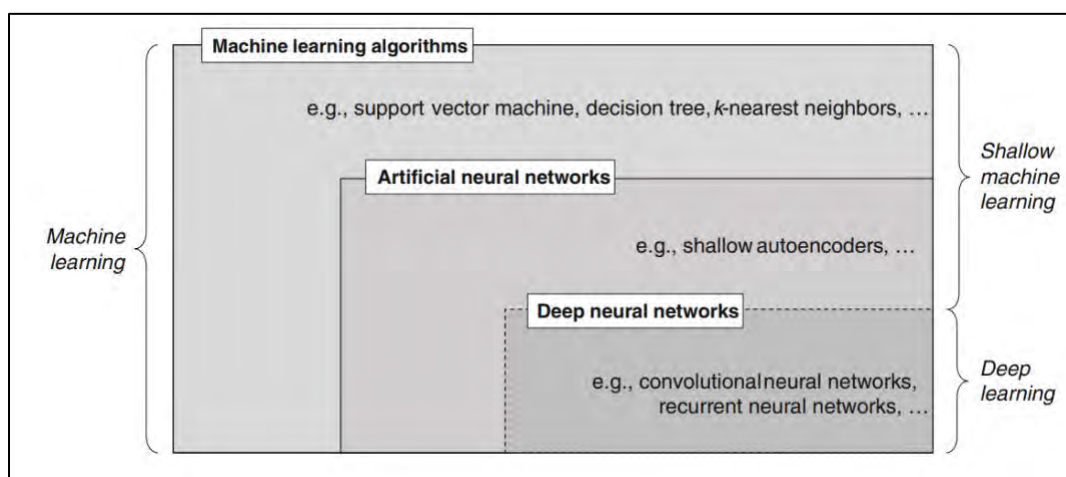
<sup>78</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>79</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 4.

<sup>80</sup> *Ibid.*

supervised, unsupervised, and reinforcement.<sup>81</sup> Additional learning methods exist, but this study will limit its scope to these three methods.

Currently, this subfield is the most prominent application of AI.<sup>82</sup> It is very popular and is used in many domains, such as medicine, automatic automobiles, marketing, networking, energy, etc.<sup>83</sup> The ML subfield is so large that it is usually divided into the following sub-subfields: shallow ML, which includes ML and Artificial Neural Networks, and Deep Learning, as per Figure 2.9.<sup>84</sup> These sub-subfields apply the same ML learning methods: supervised, unsupervised, and reinforcement.<sup>85</sup>



**Figure 2.9 – Machine Learning and Common Algorithms**

Source: Janiesch, "Machine Learning and Deep Learning," 687.

### *Artificial Neural Network (ANN)*

ANN is inspired by biological methods to process information: the brain and, more specifically, the neural network.<sup>86</sup> ANN is based on the principle of information processing in biological systems.<sup>87</sup> It uses a mathematical representation of neural networks, including neurons and weights, to adjust the strength of signals to train AI models. ANN is a method to train the model to give the most appropriate output to solve a problem. ANN has an input layer receiving

<sup>81</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>82</sup> Daniel Araya, *Artificial Intelligence for Defence and Security - Special Report* (Waterloo: 2022), 7.

<sup>83</sup> Ahmed, "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence," 43.

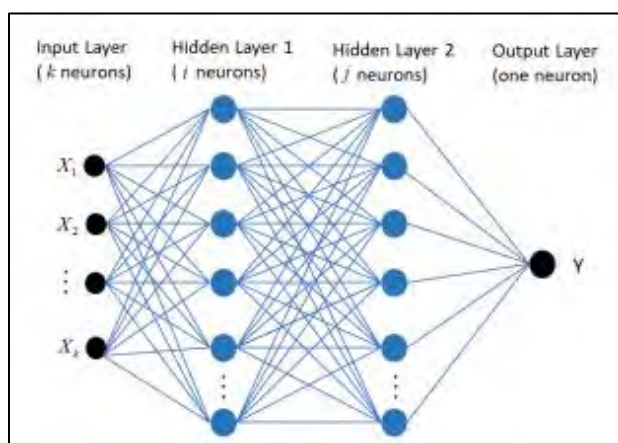
<sup>84</sup> Christian Janiesch, Patrick Zschech and Kai Heinrich, "Machine Learning and Deep Learning," *Electronic Markets* 31, no. 3 (April 8, 2021), 685. doi:10.1007/s12525-021-00475-2. <https://doi-org.cfc.idm.oclc.org/10.1007/s12525-021-00475-2>.

<sup>85</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>86</sup> Janiesch, "Machine Learning and Deep Learning," 685.

<sup>87</sup> *Ibid.*

the data, for instance, an image, and an output layer producing the results.<sup>88</sup> It could have zero or more hidden layers between these two main layers.<sup>89</sup> These hidden layers are responsible for “learning a non-linear mapping between the input and output.”<sup>90</sup> Figure 2.10 shows an ANN architecture with an input, output, and two hidden layers. It is essential to understand that each ANN architecture has only one input layer and one output layer.<sup>91</sup> The number of neurons on the input layer is the same as the input variables in the data the model will process.<sup>92</sup> The same logic applies to the output layer; the number of neurons is decided with the expected number of outputs associated with each input.<sup>93</sup> The hidden layers and neurons per hidden layer are hyperparameters, and their number is decided with optimizer algorithms as explained in the previous section in Layer 3 – Model / Intelligence Layer.<sup>94</sup>



**Figure 2.10 – ANN and DL Architecture Example**

Source: Longbin Zhang et al., *Ankle Joint Torque Estimation using an EMG-Driven Neuromusculoskeletal Model and an Artificial Neural Network Model*, Vol. 18 Institute of Electrical and Electronics Engineers (IEEE), (2021), 4.

### *Deep Learning (DL)*

Some literature separates ANN from Deep Learning (DL). The distinction between ANN and DL is very blurry. These two subfields are usually distinguished by the number of hidden layers. Once there is more than one hidden layer, the ANN becomes a DL.<sup>95</sup> In contrast to simple

<sup>88</sup> *Ibid.*

<sup>89</sup> *Ibid.*

<sup>90</sup> *Ibid.*

<sup>91</sup> "Hidden Layers of A Neural Network," last modified April 22, accessed March 22, 2024, <https://community.ibm.com/community/user/ai-datascience/blogs/moloy-de1/2023/04/16/hidden-layers-of-a-neural-network>.

<sup>92</sup> *Ibid.*

<sup>93</sup> *Ibid.*

<sup>94</sup> *Ibid.*

<sup>95</sup> Janiesch, "Machine Learning and Deep Learning," 687.

ANN, DL contains advanced neurons.<sup>96</sup> DL networks are fed raw input data and automatically discover representations needed for their learning.<sup>97</sup>

The DL subfield is very popular and has seen more advancements lately. This subfield is particularly useful in domains with large and high-dimensional data, such as text, image, video, speech, and audio data.<sup>98</sup> DL needs a large amount of data to be efficient, and it is very performant when learning with large datasets.<sup>99</sup> Otherwise, it will not be efficient and shallow ML could produce better results.<sup>100</sup> DL is often considered a further evolved subset of ML.<sup>101</sup> Nevertheless, it is important to highlight that DL is not necessarily more efficient than shallow ML.<sup>102</sup>

### *Bio-inspired Computation (BIC)*

The Bio-inspired Computation (BIC) AI subfield is relevant and commonly used in cybersecurity.<sup>103</sup> Nature often provides solutions to complex problems humans must solve. The concept of biomimicry is imitating systems found in nature to address complex problems.<sup>104</sup> The reason is that nature offers simple and sustainable solutions for many challenges. In the past, many humans have observed nature, animal behaviour and characteristics, and biology for inspiration to find solutions to complex problems.<sup>105</sup> Biomimicry has been used in many domains, such as aeronautical, where the Wright brothers studied birds to design their planes.<sup>106</sup> This method has entered the AI domain to become an AI subfield named bio-inspired computation.<sup>107</sup> The most popular bio-inspired computations are the “genetic algorithms (GA), evolution strategies (ES), ant colony optimization (ACO), particle swarm optimization (PSO), and artificial immune systems (AIS).”<sup>108</sup>

### *Natural Language Processing (NLP)*

Natural Language Processing (NLP) is a popular AI subfield since it captures almost 1/5 of the market share and solutions.<sup>109</sup> This AI subfield specializes in linguistic tasks such as

---

<sup>96</sup> *Ibid.*

<sup>97</sup> *Ibid.*

<sup>98</sup> *Ibid.*, 688.

<sup>99</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>100</sup> Janiesch, "Machine Learning and Deep Learning," 688.

<sup>101</sup> "What is an AI Model?"

<sup>102</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 4.

<sup>103</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>104</sup> Abdullah Alabdulatif and Navod Neranjan Thilakarathne, *Bio-Inspired Internet of Things: Current Status, Benefits, Challenges, and Future Directions*, Vol. 8MDPI AG, (2023), 6.

<sup>105</sup> *Ibid.*

<sup>106</sup> *Ibid.*

<sup>107</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>108</sup> *Ibid.*

<sup>109</sup> Chakraborty, *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World*, 63.

automatically analyzing and obtaining meaningful context from a text.<sup>110</sup> Therefore, NLP can understand, interpret, and generate high-quality human language and is mostly used as a chatbot.<sup>111</sup> In recent years, NLP has been significantly developed and improved due to the availability of large datasets.<sup>112</sup> This subfield will contribute to reducing the gap between humans and machines with its ability to understand human language.<sup>113</sup> Other AI subfields exist, such as robotics, computer vision, or fuzzy logic; however, for the scope of this study, the ones described above will be sufficient. The next section will describe AI model architectures.

## AI Models Architectures

This section will describe the most common AI models used in the industry, such as discriminative and generative models. It is critical to correctly identify the tasks and output required to select the appropriate model.

### *Discriminative Models*

Discriminative models are used to identify a decision boundary between classes in datasets.<sup>114</sup> They use conditional probability ( $p(x|y)$ ), which does not allow any assumptions about data. These models are useful for supervised machine learning methods.<sup>115</sup> Their main advantage is that they are robust with outliers. Finally, they require less computing power than generative models. Discriminative models use two sub-models: regression and classification.

Regression models are useful in predicting a continuous value, such as price, age, size, or time.<sup>116</sup> They determine the relationship between input ( $x$ ) and output ( $y$ ). The model will provide the corresponding output ( $y$ ) when it receives the input ( $x$ ).<sup>117</sup> Regression models are used to predict a numeric value.<sup>118</sup> Classification models will predict *discrete* values. These models are good at classifying things based on attributes. There are two types of classification models: binary and multiclassification.<sup>119</sup> The binary classification classifies data between two categories, such as “yes” or “no,” “accept” or “reject.” In contrast, multiclass classification classifies data into multiple categories, such as “product A, B, C and D” recommendations.<sup>120</sup> This model is used in many domains, such as in diagnostic image classification in radiology<sup>121</sup> or

---

<sup>110</sup> *Ibid.*

<sup>111</sup> *Ibid.*

<sup>112</sup> *Ibid.*

<sup>113</sup> *Ibid.*, 68.

<sup>114</sup> "Generative Vs. Discriminative Machine Learning Models," last modified January 2, accessed February 28, 2024, <https://www.unite.ai/generative-vs-discriminative-machine-learning-models/>.

<sup>115</sup> *Ibid.*

<sup>116</sup> "What is an AI Model?"

<sup>117</sup> *Ibid.*

<sup>118</sup> "What is Machine Learning?"

<sup>119</sup> *Ibid.*

<sup>120</sup> "What is an AI Model?"

<sup>121</sup> "What is an AI Model?"

cybersecurity. In recent years, there has been a shift in the focus of researchers from discriminative to generative.<sup>122</sup>

### *Generative Models*

Generative models, more commonly referred to as Generative (Gen) AI, differentiate themselves from discriminative models capable of recognizing and classifying patterns among data by creating new content, such as text, audio, digital images, videos and even software codes.<sup>123</sup> Generative models generate novel responses instead of pre-programmed answers.<sup>124</sup> They use unsupervised and semi-supervised learning methods.<sup>125</sup> Contrary to discriminative models that use *conditional probability* ( $p(x|y)$ ), generative models aim to predict *joint probability* ( $p(x,y)$ ) for data points emerging in a specific space.<sup>126</sup> Presently, there exist several classes or techniques of Generative AI, such as Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), and Transformer Models.<sup>127</sup> Understanding the difference between discriminative and generative models will be useful when creating AI models for solving specific problems and choosing between these approaches.

### Machine Learning Algorithms

As mentioned, ML uses three learning methods: supervised, unsupervised, and reinforcement.<sup>128</sup> These training methods are accomplished through algorithms. This section will describe the most common ML algorithm training.

### *Supervised Learning Algorithms*

The supervised learning method uses data that have been labelled. The datasets used for this type of learning contain examples with features for the input and labelled answers or target values for the output.<sup>129</sup> In other words, the input (x) is paired with output (y) data in the training set that can calibrate the model's parameters.<sup>130</sup> The supervised learning method is often used for

---

<sup>122</sup> Staphord Bengesi et al., *Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers* (United States, Ithaca: Cornell University Library arXiv.org, 2022), 2.

<sup>123</sup> Canadian Centre for Cyber Security, *Generative Artificial Intelligence - ITSAP.00.041* (Communications Security Establishment, July). <https://www.cyber.gc.ca/en/guidance/generative-artificial-intelligence-ai-itsap00041>.

<sup>124</sup> Chakraborty, *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World*, 34.

<sup>125</sup> "What is an AI Model?"

<sup>126</sup> *Ibid.*

<sup>127</sup> Chakraborty, *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World*, 38.

<sup>128</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>129</sup> Janiesch, "Machine Learning and Deep Learning," 687.

<sup>130</sup> *Ibid.*

regression and classification problems.<sup>131</sup> A common application of the supervised learning method in cyber security is the intrusion detection system (IDS), explained in Chapter 4. IDS classifies network traffic between “benign” and “malicious” or even in more categories by identifying the attack types.<sup>132</sup> This learning method is very efficient; however, it is very time-consuming since the data must be previously labelled before the training. The main supervised learning algorithms are Linear Regression, Logistic Regression, Decision Trees, Random Forest, and Support Vector Machine (SVM), which are described in Table 2.3.<sup>133</sup>

**Table 2.3 – Common Machine Learning Supervised Learning Algorithms**

<b>Algorithm Name</b>	<b>Description</b>
<b>Linear Regression</b>	This algorithm is used for regression tasks to predict an output obtained with an input feature.
<b>Logistic Regression</b>	This algorithm is frequently used for binary classification problems but can also be used for multiclass problems. It is built with a logit (log-odds) function and uses a sigmoid function to calculate the probability between 0 or 1 or ‘true’ or ‘false,’ or ‘benign’ or ‘malicious.’ <sup>134</sup>
<b>Decision Trees</b>	This algorithm can be applied to regression and classification problems. It classifies the data into small groups until these groups cannot be divided further. The algorithm uses the nodes and leaves of the tree. The nodes are decisions about data, and the leaves are the data categorized.
<b>Random Forest</b>	This algorithm is a combination of multiple decision trees. It will choose random observations and features.
<b>Support Vector Machine (SVM)</b>	This algorithm is often applied for classification tasks when the data is not easily separable. It draws the best decision boundary between data to separate different categories. <sup>135</sup>
<b>K-Nearest Neighbors (KNN)</b>	Another common algorithm used for supervised learning is a very powerful one that can be used for regression and classification tasks. This algorithm assesses the distribution of data points and separates them into classes. The algorithm calculates the distance between

<sup>131</sup> Hind Khoulimi, Mohamed Lahby and Othman Benammar, "Chapter 2 an Overview of Explainable Artificial Intelligence for Cyber Security," in *Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence*, Vol. 1025 (Warsaw: Springer, 2022), 44.

<sup>132</sup> *Ibid.*

<sup>133</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>134</sup> "Generative Vs. Discriminative Machine Learning Models."

<sup>135</sup> *Ibid.*

	two points: the greater the distance, the fewer similarities they have. <sup>136</sup>
--	--

Source: "Generative Vs. Discriminative Machine Learning Models."

### *Unsupervised Learning Algorithms*

Contrary to the supervised method, where the input (x) is paired with an output (y), the unsupervised method learns by finding common data properties, known as clustering.<sup>137</sup> The model will learn to identify meaningful patterns in the data provided.<sup>138</sup> A main difference with the supervised method is that unsupervised learning does not use labelled datasets but only unlabelled data.<sup>139</sup> In the IDS examples, the model will discover that the traffic in the network with normal behaviour is forming a denser cluster than that with abnormal behaviour that will appear as an outlier in the dataset.<sup>140</sup> The most common unsupervised algorithms for shallow ML are K-means clustering, k-nearest neighbors, and Principal Component Analysis (PCA).<sup>141</sup>

### *Artificial Neural Networks and Deep Learning Algorithms*

The ANN and DL AI subfields use the unsupervised learning method.<sup>142</sup> ANN and DL use unsupervised algorithms to select features automatically. The most common algorithms used for DL are "Feed Forward Networks (FNN), Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Deep Belief Networks (DBNs), Stacked Autoencoders (SAE), Generative Adversarial Networks (GANs), Restricted Boltzmann Machines (RBMs), and Ensemble of DL Networks (EDLNs)."<sup>143</sup> Table 2.5 below defines some of these common algorithms.

**Table 2.4 – Common Deep Learning Algorithms**

Algorithms Name	Description
Convolutional Neural Networks (CNNs)	This algorithm is often used for computer vision, speech recognition, images, and videos.
Recurrent Neural Networks (RNN)	This algorithm is often employed for sequential data such as time series or natural language.
Autoencoder	Autoencoders distinguish themselves with the way they train the model: they corrupt the

<sup>136</sup> "What is a KNN (K-Nearest Neighbors)?" Accessed May 6, 2024. <https://www.unite.ai/what-is-k-nearest-neighbors/>.

<sup>137</sup> Janiesch, "Machine Learning and Deep Learning," 687.

<sup>138</sup> "What is Machine Learning?"

<sup>139</sup> Khoulimi, "Chapter 2 an Overview of Explainable Artificial Intelligence for Cyber Security," in , , 44

<sup>140</sup> Khoulimi, "Chapter 2 an Overview of Explainable Artificial Intelligence for Cyber Security," 44.

<sup>141</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 5.

<sup>142</sup> *Ibid.*

<sup>143</sup> *Ibid.*

	training data to encourage the model to learn the key features of the data. <sup>144</sup> The data is compressed and encoded, then decoded and reconstructed to reproduce data as closely as possible to the original data. <sup>145</sup>
Generative Adversarial Networks (GAN)	GAN is proposing an architecture where two models are combined: the generator model and the discriminator model. <sup>146</sup> The generator model will create new data from the patterns learned during the training on the training datasets. <sup>147</sup> Meanwhile, the discriminator model will analyze the input, usually images, and must classify them as fake or real. <sup>148</sup> Therefore, the ML model (discriminator) is trained to find flaws in the output generated by another ML model (generator). <sup>149</sup>

Sources: Janiesch, "Machine Learning and Deep Learning," 690.

"What is an Autoencoder?"

### *Reinforcement Learning Algorithm*

This technique does not provide input (x) and output (y) pairs. The ML model has the current state, specific goal, list of allowable actions and constraints for the outcome.<sup>150</sup> The model will learn to get rewards using the principle of trial and error.<sup>151</sup> The reinforcement is done through a feedback mechanism completely automated.<sup>152</sup> This method is very successful in closed-world environments such as games.<sup>153</sup> It is also used to train robots.<sup>154</sup> The most common

<sup>144</sup> "What is an Autoencoder?" last modified September 20, 2024, accessed March 22, 2024, <https://www.unite.ai/what-is-an-autoencoder/>.

<sup>145</sup> *Ibid.*

<sup>146</sup> "What is a Generative Adversarial Network (GAN)?" last modified October 5, accessed March 26, 2024, <https://www.unite.ai/what-is-a-generative-adversarial-network-gan/>.

<sup>147</sup> *Ibid.*

<sup>148</sup> *Ibid.*

<sup>149</sup> Khalifa Al-Dosari, Noora Fetais and Murat Kucukvar, "Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges," *Cybernetics and Systems* 55, no. 2 (Aug 12, 2022), 5. doi:10.1080/01969722.2022.2112539. <https://doi.org/10.1080/01969722.2022.2112539>.

<sup>150</sup> Janiesch, "Machine Learning and Deep Learning," 687.

<sup>151</sup> *Ibid.*

<sup>152</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 4.

<sup>153</sup> Janiesch, "Machine Learning and Deep Learning," 687.

<sup>154</sup> "What is Machine Learning?"

algorithms or methods for reinforcement learning are Q-learning, Deep Q Network, or Markov Decision Process.<sup>155</sup>

## Summary

Chapter 2 aimed to explain key AI concepts so the reader could better understand the recommendations in the next two chapters. Section 2.1 described layers that formed AI models. As mentioned, the names of layers can vary in the literature; however, the most common four layers are the data layer, machine learning framework and algorithms layer, model/intelligence layer, and application layer. Section 2.2 proposed classifications and definitions between AI subfields, AI model architectures, and machine learning algorithms.

---

<sup>155</sup> Johannes Czech "Distributed Methods for Reinforcement Learning Survey," In *Reinforcement Learning Algorithms: Analysis and Applications*, edited by Belousove, Boris, Hany Abdulsamad, Pascal Klink, Simone Parisi and Jan Peters, (Springer Nature Switzerland, 2021), 154.

## CHAPTER 3 – LEVERAGING AI

Even if leveraging AI models has multiple advantages, this does not come without risks. As a result, any private or public organizations that aspire to leverage AI must be aware of the risks, regulations, and compliance related to AI. This chapter intends to demonstrate that the CAF Cyber Command leadership and staff must be cognizant that leveraging AI models brings additional responsibilities and risks and must be prepared to defend them against adversaries. Accordingly, this chapter will be divided into four sections. The first section will propose a 10-step framework to build AI models. The second section will discuss AI models' risks and vulnerabilities and recommend methods to counter these risks and harden the defence of deployed AI models. The third section will detail the Government of Canada's regulations and compliance regarding AI. Finally, the fourth section will present an emerging research field that aims to improve the transparency and trust of AI models.

### Section 3.1 – Framework for Creating AI Models

#### In-House Development versus Commercial-Off-the-Shelf Product

Before deciding to leverage AI, an important decision is whether the CAF Cyber Command should develop its own model or buy a commercial-off-the-shelf (COTS) product. The main advantage of fully developing its AI model solution is having full control over the model; however, this control is not free.<sup>156</sup> If CAF decides to build its model, it must expect important effort, time, and financial resources dedicated to building and maintaining it.<sup>157</sup> Developing the model also requires a high level of expertise to make challenging decisions about the choice of data, features, architecture, training algorithms, parameters, hyperparameters, etc.<sup>158</sup>

The opposite of in-house development is buying or renting a commercial-off-the-shelf (COTS) solution. Specifically in cybersecurity, there are two downsides to using COTS solutions: limited scope and lack of transparency.<sup>159</sup> Some COTS solutions cannot be easily modified to meet specific organizational goals.<sup>160</sup> Also, these AI models may not be trained on the organization's data.<sup>161</sup> The advertised performance of COTS solutions is often based on the vendor environment where the model was built. The actual performance in the buyer environment might be different.<sup>162</sup> Therefore, it might be difficult to integrate the model into the organizational environment because it is different from the vendor environment in which the

---

<sup>156</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 18.

<sup>157</sup> *Ibid.*

<sup>158</sup> *Ibid.*

<sup>159</sup> *Ibid.*, 20.

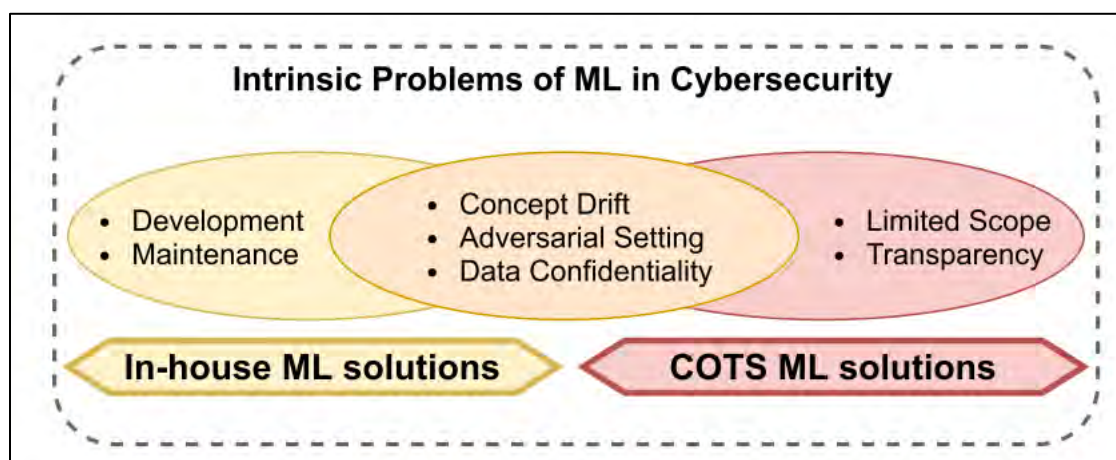
<sup>160</sup> *Ibid.*, 19.

<sup>161</sup> *Ibid.*

<sup>162</sup> *Ibid.*

model was built. Therefore, the model's efficiency might be reduced.<sup>163</sup> For instance, in cyber security, some AI models can detect malicious activities that can be applied in every environment, such as a compromised website.<sup>164</sup> However, an AI model might not detect activities such as malicious behaviour since these behaviours would be specific to a network.<sup>165</sup> That is why most COST solutions will focus on shallow ML with unsupervised training algorithms for anomaly detection.<sup>166</sup> An additional problem of using COST solutions is their lack of transparency.<sup>167</sup>

For any project that requires leveraging AI, CAF Cyber Command will have to decide between an in-house solution, a COTS solution, or a combination of these two approaches. Each option has advantages and disadvantages, which are described in Figure 3-1. The disadvantage of in-house solutions is the resource requirement for their development and maintenance, whereas the COTS solution has a limited scope and transparency issues.



**Figure 3. 1 – Chapter 1 Disadvantages of In-house ML Solutions and COTS**

Source: Apruzzese, "The Role of Machine Learning in Cybersecurity," *Digital Threats: Research and Practice*," 16.

If the CAF Cyber Command decides to build its own AI models to bolster its DCOs, it is recommended that it follow a framework, which will be described below.

### 10-Step Framework

This section will propose a 10-step framework for creating AI models. This framework is very similar to other frameworks used in the industry. However, it is purposely intended to be

<sup>163</sup> *Ibid.*

<sup>164</sup> *Ibid.*

<sup>165</sup> *Ibid.*

<sup>166</sup> *Ibid.*

<sup>167</sup> *Ibid.*

large and slightly vague since the requirements to build an AI model can change rapidly. It does not explicitly provide all the details to build an AI model, but it provides the big picture of what needs to be accomplished and in which order. Furthermore, this framework seeks to include the Government of Canada's AI compliance and be aligned with *The DND and CAF AI Strategy* by considering its LoEs.<sup>168</sup> This proposed framework should be used to build an AI model for DCO's purposes; however, it is believed that it could be applied to any domain. Table 3.1 summarizes the 10-step framework.

**Table 3.1 – 10-Step Framework Summary**

Steps	Description
<b>Step 1 – Framing the Problem</b>	<p>Answer the following questions:</p> <ul style="list-style-type: none"> <li>• What is the problem that needs to be solved?</li> <li>• What is the context?</li> <li>• What are the tasks that need to be accomplished?</li> <li>• What are the objectives to achieve?</li> <li>• What kind of output is needed to solve the problem?</li> <li>• Is there a requirement to understand the internal workings of the model for transparency, accountability, and trust?</li> <li>• Does the organization have already an AI model that achieve this objective that could be used?</li> <li>• Is AI the only way to solve the problem?</li> </ul>
<b>Step 2 – Datasets</b>	This step selects and prepares datasets to train, validate and test the model.
<b>Step 2A – Raw Data Collection</b>	<ul style="list-style-type: none"> <li>• Collecting raw data to build an in-house dataset.</li> <li>• Identifying public datasets that can be used.</li> </ul>
<b>Step 2B – Identify Features and Labels</b>	<ul style="list-style-type: none"> <li>• Identifying features.</li> <li>• Labelled data (if applicable).</li> </ul>
<b>Step 2C - Preprocessing</b>	<ul style="list-style-type: none"> <li>• Cleaning the data.</li> <li>• Feature engineering.</li> </ul>
<b>Step 2D – Sampling and Split the Datasets</b>	Sampling and splitting the datasets into the training, validating, and testing sets.
<b>Step 3 – Selecting the Model</b>	This step includes three sub-steps.

<sup>168</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 32.

<b>Step 3A – Selecting the AI Subfield</b>	Selecting the most appropriate AI subfield for the problem between expert systems, shallow ML, DL, or BIC.
<b>Step 3B – Selecting the AI Model</b>	Selecting the most appropriate model architecture for the problem between discriminative and Gen AI.
<b>Step 3C – Selecting the Training Algorithms</b>	Selecting the most appropriate training algorithms between supervised, unsupervised, reinforcement, and DL algorithms.
<b>Step 4 – Infrastructure Requirement</b>	Identifying the infrastructure requirements for the AI model.
<b>Step 5 – Training the Model</b>	Train the model with the training set.
<b>Step 6 – Validating the Model</b>	<ul style="list-style-type: none"> <li>• Validate the model with the validation set.</li> <li>• For supervised algorithms, check for overfitting and underfitting.</li> <li>• Adjust parameters and hyperparameters.</li> <li>• Complete the Algorithmic Impact Assessment</li> <li>• Complete the Gender-Based Analysis Plus (GBA+)</li> </ul>
<b>Step 7 – Testing the Model</b>	Test the model with the testing set.
<b>Step 8 – Deploying the Model</b>	<ul style="list-style-type: none"> <li>• Deployment of the model in the production environment.</li> <li>• Publish policies related to the model.</li> </ul>
<b>Step 9 – Monitoring and Maintaining the Model</b>	Monitor, retrain, and update the model regularly.
<b>Step 10 – Defending the Model</b>	Defend the model against adversarial machine learning.

Before starting this process, it is strongly recommended that all steps in the proposed framework be documented. This will help justify the project and meet the requirements detailed later in this chapter.<sup>169</sup>

### Step 1 – Framing the Problem

Identifying the problem to be solved is the first step in most AI model creation frameworks in the industry. Understanding what is the problem the CAF Cyber Command leadership wants to solve is probably the most important step in any AI model. Otherwise, the model will not accomplish the initial intent and might cost significant resources regarding time, human

<sup>169</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 18.

expertise, and money. As mentioned in Chapter 1, the *PFEC* insisted on the importance of framing the problem properly before leveraging AI.<sup>170</sup> Several questions should be answered to correctly grasp the intent, hence the name of this step, “Framing the Problem.” Here is a recommended series of questions that should be answered and documented at this stage of the project:

- What is the problem or the issue that needs to be solved?
- What is the context?
- What are the tasks that need to be accomplished?
- What is the objective(s)? Knowing the objective will help decide whether discriminative or generative AI is required to solve the problem.
- What kind of output is needed to solve the problem? Knowing the expected output, such as text, image, video, etc., will be useful when selecting the AI subfield, model architecture, and algorithms.
- Is there a requirement to understand the internal workings of the AI model for transparency, accountability, and trust? This question is important because it will help select the type of AI model between black-box (non-interpretable) or white-box (interpretable) AI models.
- Does the organization have already an AI model that achieve this objective that could be used? It is necessary to verify if the solution already exists to avoid waste of resources and multiplication of efforts between organizations.
- Is AI the only way to solve the problem? AI, and specifically ML, is not the solution to all problems.<sup>171</sup> It is important to avoid the mistake of using AI for the sake of AI. AI is not always the best option for achieving an objective.<sup>172</sup>

Answering these questions properly is important to determine if AI can solve the problem. It will improve efficiency and avoid building models that are too complex and require too many resources for a problem requiring a simpler AI model. For instance, a problem might need only a discriminative model, not a generative one, which could be much costlier and more complex to build. Understanding the context in which the AI model will be used is important to help select the right approach later. Moreover, at this stage, it is recommended to start involving legal services, which will become useful to ensure that all regulations and compliance will be followed.<sup>173</sup> Finally, this step should end by knowing the problem to be solved, the objectives to be accomplished, and what kind of output is required.

---

<sup>170</sup> National Defence, *Pan-Domain Force Employment Concept - Prevailing in a Dangerous World*, 26.

<sup>171</sup> Veda C. Storey et al., "Viewpoint - Explainable AI," *Communications of the ACM* 65, no. 4 (April, 2022), , 28. doi:10.1145/3490699. <https://doi.org/10.1371/journal.pone.0231166>.

<sup>172</sup> Ventre, *Artificial Intelligence, Cybersecurity and Cyber Defence*, 166.

<sup>173</sup> "Algorithmic Impact Assessment Tool," last modified April, 25, accessed March 15, 2024, <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

## Step 2 – Selecting and Preparing Datasets

The second step involves selecting or building datasets to train, validate, and test the AI model. This stage could be considered the most important since the output quality will likely depend on the data quality used to train the model.<sup>174</sup> Moreover, most of the time spent building the model will be allocated at this stage. Constructing an in-house dataset could be split into four steps: collecting the raw data, preprocessing, sampling the data, and splitting the data.<sup>175</sup>

### *Step 2A – Raw Data*

The first step is collecting raw data to build an in-house dataset. At this stage, the dataset's size and quality will be decided.<sup>176</sup> Figure 3.2 below shows examples of different dataset sizes.

Data set	Size (number of examples)
Iris flower data set	150 (total set)
MovieLens (the 20M data set)	20,000,263 (total set)
Google Gmail SmartReply	238,000,000 (training set)
Google Books Ngram	468,000,000,000 (total set)
Google Translate	trillions

**Figure 3.2 – Example of Dataset's Size**  
Source: "Data Prep - Construct Your Dataset."

Reliability, feature representation, and minimizing skew are important considerations when improving the quality of the dataset.<sup>177</sup> The Canadian Centre for Cyber Security mentions that the data used for training the model "should be complete, diverse, and accurate."<sup>178</sup> These criteria will ensure that all patterns are discovered accurately and provide reliable results.<sup>179</sup> It is important to collect relevant data. The datasets must be of sufficient quality and quantity so the model can perform well.<sup>180</sup>

<sup>174</sup> "Data Prep - Construct Your Dataset," last modified July 18, accessed April 3, 2024, <https://developers.google.com/machine-learning/data-prep/construct/construct-intro>.

<sup>175</sup> *Ibid.*

<sup>176</sup> *Ibid.*

<sup>177</sup> *Ibid.*

<sup>178</sup> Canadian Centre for Cyber Security, *Artificial Intelligence - ITSAP.00.040*

<sup>179</sup> *Ibid.*

<sup>180</sup> "What is an AI Model?"

### *Step 2B – Identify Feature and Label Sources*

Once the data are collected, it is required to identify the features and label the data. Supervised algorithms will require labelled data to complete their training.<sup>181</sup> There are two types of labelling: direct and derived labels.<sup>182</sup> The direct label is what explicitly what prediction is required, for instance, “malicious email,” whereas a derived label is a reliable indicator that does not directly measure the prediction, for instance, “email containing a file.”<sup>183</sup> Labelling can be done manually with human raters; however, this is an expensive process, and mistakes are possible.<sup>184</sup>

### *Step 2C – Preprocessing*

Step 2C is about preprocessing the datasets, which includes data cleaning and feature engineering.<sup>185</sup> The first step, data cleaning, requires close collaboration between data scientists and domain specialists, who will use the AI model, such as cyber operators, in the context of cyber operations.<sup>186</sup> This step also requires a high level of understanding of the context of the AI model.<sup>187</sup> Data cleaning involves removing duplicate, irrelevant, or incorrect data, joining data from different files, and addressing missing data.<sup>188</sup> The second step, feature engineering, is about data normalization, transformation, feature selection, dimensionality reduction, or data type conversion to meet the algorithm's requirements.<sup>189</sup> At the end of Step 2C, datasets will be selected and cleaned, and the features engineering will be completed.

### *Step 2D – Sampling and Splitting the Data*

The fourth step requires sampling and splitting the selected datasets into three sets: training, validating, and testing.<sup>190</sup> It is important to note that each example from the dataset can only be in one category. In other words, an example cannot be in both the training and validation datasets. Table 3.2 defines these three categories. The training set will be applied to train the model. Then, the validation set will fine-tune the model hyperparameters. Finally, the testing set will complete the final evaluation of the model.

---

<sup>181</sup> Zhang, "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities," 1042.

<sup>182</sup> "Data Prep - Construct Your Dataset."

<sup>183</sup> *Ibid.*

<sup>184</sup> *Ibid.*

<sup>185</sup> André Pfob, Sheng-Chieh Lu and Chris Sidey-Gibbons, "Machine Learning in Medicine: A Practical Introduction to Techniques for Data Pre-Processing, Hyperparameter Tuning, and Model Comparison," *BMC Medical Research Methodology* 22, no. 1 (-11-01, 2022), 2. doi:10.1186/s12874-022-01758-8. <https://doi.org/10.1186/s12874-022-01758-8>.

<sup>186</sup> *Ibid.*

<sup>187</sup> *Ibid.*

<sup>188</sup> *Ibid.*

<sup>189</sup> *Ibid.*

<sup>190</sup> "Data Prep - Construct Your Dataset."

**Table 3.2 – Datasets Splitting Categories and Definitions**

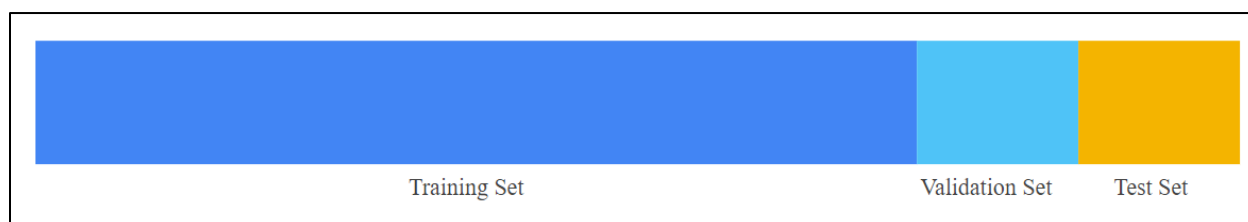
<b>Dataset Splitting Categories</b>	<b>Definitions</b>
Training Set	A sample of the original dataset is used to train and fit the model.
Validation Set	After the model is trained, a validation set will be used for initial evaluation before using the testing set. The validation set can be used for several iterations, allowing tuning the model hyperparameters. Since the validation datasets are used several times, the evaluation eventually becomes biased.
Testing Set	A sample of the original dataset is used against the model for an unbiased evaluation.

Sources: "What is the Difference between Test and Validation Datasets?"

"Data Prep - Construct Your Dataset."

"Machine Learning Glossary - Training Set."

Figure 3.3 shows a dataset split into the training, validation, and testing sets.

**Figure 3.3 – Slicing a Dataset into Three Sets for Training, Validation, and Testing**

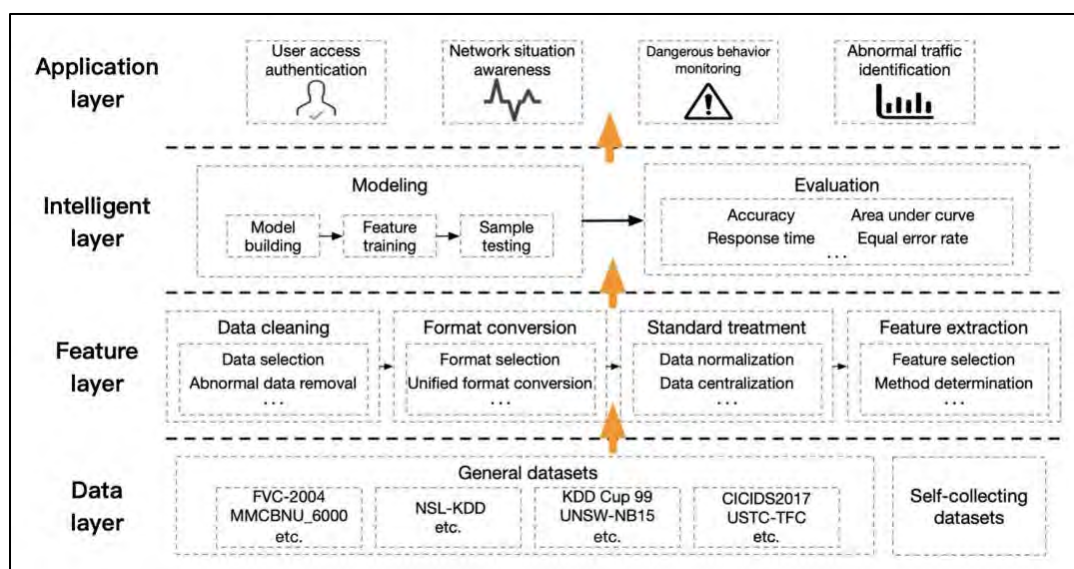
Source: "Machine Learning Crash Course - Validation Set: Another Partition."

### *Public Datasets*

In the cyber security domain, multiple available public datasets can be used to train AI models. For instance, the CSE-CIC-IDS2018 can be found on Open Data on AWS.<sup>191</sup> The training datasets must be selected accordingly for the problem and expected output identified during Step 1. As illustrated in Figure 3.4 below, which was presented in Chapter 2, the Data Layer shows that several datasets can be selected and combined with in-house datasets to train the model. Synthetic datasets are also an alternative solution to real datasets. They are smaller and resemble real datasets. Their advantage is that they help reduce privacy concerns.<sup>192</sup>

<sup>191</sup> "A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)," last modified n.d., accessed May 4, 2024, <https://registry.opendata.aws/cse-cic-ids2018/>.

<sup>192</sup> "What is an AI Model?"



**Figure 3.4 – Model for Cyber Security**

Source: Zhang, "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities," 1042.

### *Summary of Step 2*

This step is critical since it requires selecting the right datasets to train, validate, and test the model. If the data is inadequate, it will create problems in the output, such as bias, hallucinations, and trust.<sup>193</sup>

### *Step 3 – Selecting the Model*

All explanations in Chapter 2 will become useful at this stage, in which several decisions will be required to select the best AI subfield, architecture model, and algorithms to solve the problem identified in Step 1.

#### *Step 3A – Selecting the AI Subfield*

Step 3A requires selecting the most appropriate subfield between expert systems, shallow machine learning, deep learning, and bio-inspired computing.

#### *Step 3B – Selecting the AI Model Architecture*

Step 3B requires selecting the most appropriate architecture between discriminative or generative AI. The types of problems and the expected output identified during Step 1 will help

<sup>193</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 5.

select the most appropriate architectures. It is important to select the simplest architectures to achieve the objective. In other words, even if generative AI is popular, it is not always the right solution.

### *Step 3C – Selecting the Training Algorithms*

Step 3C requires choosing the training algorithms between supervised, unsupervised, reinforcement or deep learning algorithms. Several considerations will help to select the right training algorithms. For instance, if the datasets cannot be labelled, unsupervised, reinforcement, or DL algorithms will be more appropriate. For labelled datasets, supervised algorithms are the most appropriate. Another example would be if there is a strict requirement for the transparency and interpretability of the model, it is recommended to use supervised algorithms.

### *Step 4 – Infrastructure Requirements*

Step 4 requires identifying the infrastructure requirements for the AI models, such as storage, computational power, level of security classification, cloud computing, etc. Complex AI models will demand significant computational power and bigger datasets and will be more expensive to store.<sup>194</sup>

### *Step 5 – Training the Model*

At this stage, the model starts its training. Models are trained to predict a desired output based on a provided input. Training the model involves showing data to gradually learn the patterns and relationships between features and labels or just between features.<sup>195</sup> Once the training is completed, the trained model must predict unlabeled examples. This process is called inference.<sup>196</sup>

### *Step 6 – Validating the Model*

Once the training has been completed, validating the model and making adjustments is important. As explained in Chapter 2, the validation set will be used at this stage.<sup>197</sup> The validation set will be used to verify if the outputs provided by the models are accurate.<sup>198</sup> Multiple aspects must be validated, such as the model's performance, computational resources, and interpretability.<sup>199</sup> Since there are endless options for AI model architecture, algorithms,

---

<sup>194</sup> "What is Artificial Intelligence (AI)?"

<sup>195</sup> "Machine Learning Crash Course – Framing."

<sup>196</sup> *Ibid.*

<sup>197</sup> "Machine Learning Architecture: The Code Components," last modified n.d., accessed March 28, 2024, <https://www.datarevenue.com/en-blog/machine-learning-project-architecture>.

<sup>198</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 4.

<sup>199</sup> Janiesch, "Machine Learning and Deep Learning," 691.

parameters, hyperparameters and training datasets, it is important to find the right balance between cost-efficiency, robustness and privacy in the model.<sup>200</sup>

### *Overfitting and Underfitting*

Overfitting and underfitting are key concepts that must be understood during the validation. Overfitting means that the model output fits exactly against the training sets.<sup>201</sup> Overfitted models will not be accurate with the testing set. They have memorized the noise and cannot generalize new data.<sup>202</sup> Labelling data will help reduce the overfitting.<sup>203</sup> Underfitting happens when the model has not trained enough against the training set or when the input variables are not enough to enable the model to determine the relationship between the input and output.<sup>204</sup> Underfitted models are considered to have bias.<sup>205</sup> Overfitting and underfitting impact only ML that uses supervised learning methods. DL models use unsupervised methods and learn by extracting what is useful from the data.<sup>206</sup> Therefore, they do not overfit or underfit.

Several iterations of validation can be used to fine-tune the model. Once the model has provided the expected output, the next step is to test it. Additionally, some compliance requirements should be done during this step. First, it would be recommended to complete the Algorithmic Impact Assessment, which will be discussed in Section 3.3 of this chapter.<sup>207</sup> Second, the Gender-based Analysis Plus (GBA+) should also be done at this stage of the projects.<sup>208</sup>

### Step 7 – Testing the Model

The difference between Steps 6 and 7 is that the validation will be used to evaluate the results after the training. In contrast, the testing is used as a final assessment once the model has been successfully validated. Figure 3.3 shows the workflow between the validation step, where several iterations can be done, and the testing step at the end.

---

<sup>200</sup> *Ibid.*

<sup>201</sup> Ahmed, "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence," 54.

<sup>202</sup> *Ibid.*

<sup>203</sup> *Ibid.*

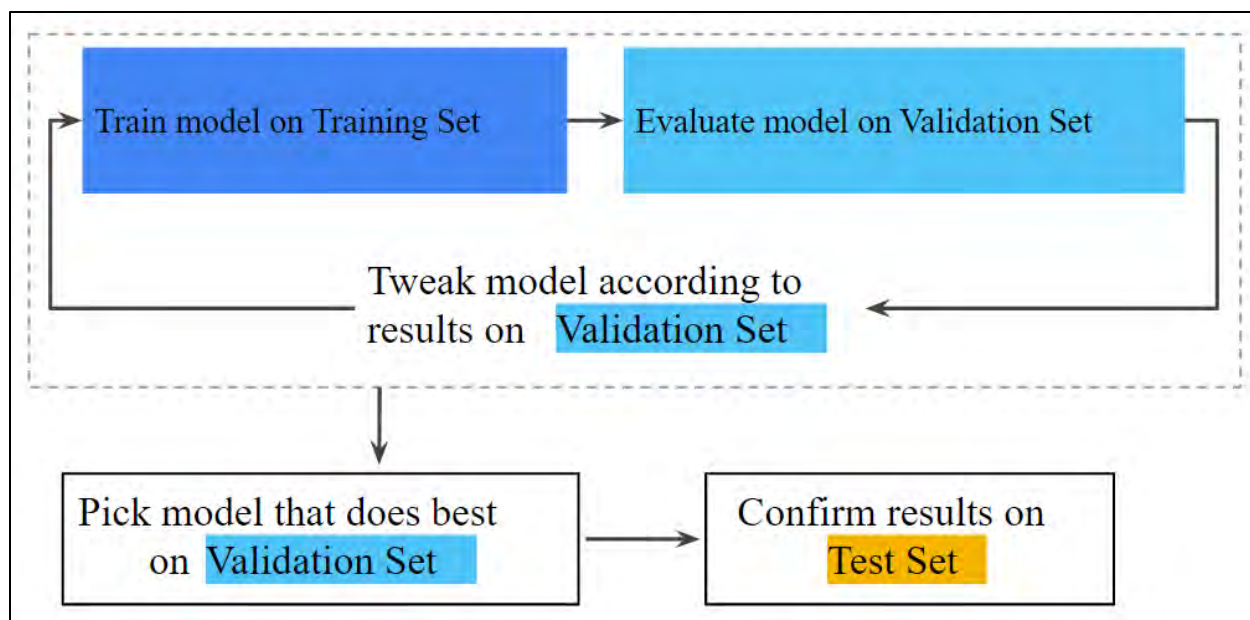
<sup>204</sup> *Ibid.*

<sup>205</sup> *Ibid.*

<sup>206</sup> *Ibid.*

<sup>207</sup> "Algorithmic Impact Assessment Tool," last modified April, 25, accessed March 15, 2024, <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.

<sup>208</sup> "Gender-Based Analysis Plus (GBA Plus)," last modified October 13, accessed April 22, 2024, <https://www.canada.ca/en/women-gender-equality/gender-based-analysis-plus.html>.



**Figure 3.5 – Validation and Testing Workflow**

Source: "Machine Learning Crash Course - Validation Set: Another Partition."

#### Step 8 – Deploying the Model

Once the model is successful after the testing phase, it is ready to be moved from the testing environment to the production environment—in other words; it is ready to be used. At this stage, other important aspects must be accomplished, such as informing and training users on how to use the model.<sup>209</sup> Policies related to the model must also be written and published.

#### Step 9 – Monitoring and Maintaining the Model

Once deployed in a production environment, AI models require constant monitoring and maintenance. This step ensures the model performs as expected and detects any divergence between actual and expected behaviour.<sup>210</sup> Also, the model might need to be trained and updated frequently with more recent datasets to reduce the effect of concept drift. The effect of concept drift happens when the model does not perform properly with the current data because they don't reflect the current trends.<sup>211</sup> This is particularly important for AI models used for cyber security since the effect of concept drift is enhanced by the changing nature of the cyber domain. Updating the training set and retraining the model is not an easy task. For instance, it is not only

<sup>209</sup> Davy Van De Sande et al., "Developing, Implementing and Governing Artificial Intelligence in Medicine: A Step-by-Step Approach to Prevent an Artificial Intelligence Winter," *BMJ Health Care Inform* 29, no. 1 (-02, 2022), 2. doi:10.1136/bmjhci-2021-100495. <https://doi.org/10.1136/bmjhci-2021-100495>.

<sup>210</sup> Mariarosaria Taddeo, Tom Mccutcheon and Luciano Floridi, "Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword," *Nature Machine Intelligence* 1, no. 12 (November 11, 2019), 560. doi:10.1038/s42256-019-0109-1. <https://doi.org/10.1038/s42256-019-0109-1>.

<sup>211</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 18.

about adding new examples reflecting the current trends in the set. Decisions must also be made about whether older examples should be kept or removed from the training set.<sup>212</sup>

### Step 10 – Defending the Model

After deploying, monitoring, and maintaining the AI model, the next step is to defend it. Criminals, advanced persistent threats (APT), and state actors will certainly target the DND and CAF AI-enabled systems. If not defended properly against adversarial machine learning, explained in detail in the next section, the AI model could become useless or even dangerous because its output will not be reliable anymore. Every time the CAF deploys an AI model, the CAF's attack surface increases. Therefore, additional resources to defend the model must be allocated.

### Summary

This 10-step framework intends to be a general guideline for creating in-house AI models. Again, it does not include all the technical details to build an ML model. Several key considerations to bear in mind when using this framework are the following: First, Step 1—Framing the Problem must be done properly; otherwise, the project might fail, and the model will not accomplish the objective. Also, it will ensure that the right subfield, architecture, and algorithms are selected to solve the problems and will reduce resources wasted. Second, Step 2 – Selecting and Preparing Datasets is crucial to the success of the AI model. Nothing in this step can be neglected. Moreover, a considerable amount of time will be spent at this stage. Third, Step 9 – Monitoring and Maintaining will ensure the model keeps performing. Finally, Step 10 – Defending the Model is critical; otherwise, instead of solving a problem, the model will create additional problems. The next section will detail the common vulnerabilities in AI models and propose solutions to ensure their defence.

## Section 3.2 – Defending the Model

It is critical to be cognizant of the risks regarding AI. Knowing the risks and implementing mitigation measures will avoid unacceptable damage to the organization. The scope of this study can neither identify all possible risks nor recommend all possible mitigation measures related to AI. Therefore, only the main risks will be discussed. They will be divided into two categories: output risks and adversarial machine learning.

### Output Risks

As mentioned in Chapter 2, the output is the *raison d'être* of each AI model. It is the component that provides the expected answer from the AI model. In other words, it is the solution to the problem identified during Step 1 – Framing the Problem, that needs to be solved.

---

<sup>212</sup> *Ibid.*

Because of its importance, the output will generate some risks, especially with Gen AI. The output can be biased, harmful, untrustful, not transparent, mistaken, or breach privacy.<sup>213</sup> Gen AI models available to the public can propagate bias and erroneous information that could significantly damage and hurt individuals and institutions, such as democracies or criminal justice systems.<sup>214</sup> Gen AI can also violate intellectual property and individual privacy rights.<sup>215</sup> AI models, in their essence, are objective and don't have prejudice; however, input and training data can be biased.<sup>216</sup>

### *Hallucinations*

Another downside of Gen AI is its risk of hallucinating. AI hallucinations are nonsensical and inaccurate outputs when the model perceives patterns among data that don't exist or cannot be observed by humans.<sup>217</sup> The main factor that causes hallucinations is incomplete or biased training data.<sup>218</sup> There are different types of hallucinations, such as "incorrect predictions, false positives and false negatives."<sup>219</sup> Hallucinations could harm the trust in Gen AI models.

### Mitigation Measures Against Output Risks

The *Canadian Centre for Cyber Security* published the cyber security guidance, *Generative Artificial Intelligence – ITSAP.00.041*, to provide advice related to Gen AI.<sup>220</sup> It emphasized the importance of Gen AI usage policies, the selection of training datasets, and the choice of security-oriented commercial tools.<sup>221</sup>

### *AI Usage Policies*

Usage policies should be provided to users for each deployed AI-enabled system. They should include the type of content that could be generated and how the organization uses technology to protect against compromises of sensitive data. The policy should also describe all processes required to use the system appropriately.<sup>222</sup> Moreover, the user must understand the

---

<sup>213</sup> "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems," last modified January 12, accessed March 17, 2024, <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

<sup>214</sup> *Ibid.*

<sup>215</sup> *Ibid.*

<sup>216</sup> Michael Haenlein and Andreas Kaplan, "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence," *California Management Review* 61, no. 4 (July 17, 2019), , 11. doi:10.1177/0008125619864925. <https://doi.org/10.1177/0008125619864925>.

<sup>217</sup> "What are AI Hallucinations?" last modified n.d., accessed March 18, 2024, <https://www.ibm.com/topics/ai-hallucinations>.

<sup>218</sup> *Ibid.*

<sup>219</sup> *Ibid.*

<sup>220</sup> Canadian Centre for Cyber Security, *Generative Artificial Intelligence - ITSAP.00.041*

<sup>221</sup> *Ibid.*

<sup>222</sup> *Ibid.*

policies.<sup>223</sup> They should be written during all steps of the 10-step framework and published during Step 8 – Deploying the Model when the AI model is put in production.

### *Training Sets*

Most output risks are related to the training sets. A robust process for selecting, validating, and verifying datasets is recommended.<sup>224</sup> Datasets must be diverse and representative to reduce inaccurate and biased content. Furthermore, there are efforts and research to reduce and even eliminate hallucinations from Gen AI. Google Cloud proposed some mitigation techniques against hallucinations. The first technique, regularization, penalizes the model when it makes extreme predictions. This technique will limit possible outcomes that a model is allowed to produce.<sup>225</sup> The second mitigation is related to the training data. It is recommended to use only data that are relevant to the task.<sup>226</sup> For instance, BMW's LLM-enhanced voice assistant is restricted to only using the internal BMW documentation about the car.<sup>227</sup> This process is called the Retrieval-Augmented Generation.<sup>228</sup> The third solution is to create a template for your AI model to follow, such as the structure of an essay.<sup>229</sup> Finally, Google Cloud suggests telling the AI model what you want and don't want through feedback.<sup>230</sup>

### *Review Outputs*

Establishing a process to review the AI model's output is suggested. Diverse stakeholders and teams across the organization should review output to check for harmful, biased, and hallucinations. Then, continuously fine-tuning or retraining the model to improve output quality is recommended.<sup>231</sup>

### *Security-Oriented Tools*

Unless built from scratch, most AI models will require leveraging vendor or third-party tools, platforms, datasets, large language models, or foundation models. Therefore, the selection process must be done carefully, and security must be a critical criterion.

---

<sup>223</sup> *Ibid.*

<sup>224</sup> *Ibid.*

<sup>225</sup> "What are AI Hallucinations?"

<sup>226</sup> *Ibid.*

<sup>227</sup> "BMW showed Off Hallucination-Free AI at CES 2024," last modified January 17, accessed March 20, 2024, <https://arstechnica.com/cars/2024/01/bmws-ai-powered-voice-assistant-at-ces-2024-sticks-to-the-facts/>.

<sup>228</sup> *Ibid.*

<sup>229</sup> "What are AI Hallucinations?"

<sup>230</sup> *Ibid.*

<sup>231</sup> Canadian Centre for Cyber Security, *Generative Artificial Intelligence - ITSAP.00.041*.

### *Corporate Classified and Sensitive Data*

It is strongly recommended that identifiable information (PII) and the organization's classified and sensitive data never be provided when selecting training and input data.<sup>232</sup> The dataset should be verified not to contain any PII.<sup>233</sup> This task should be done during Step 2A—Raw Data of the 10-step framework.

To summarize, output risks, such as hallucinations or biased output, can be reduced by establishing AI usage policies, properly selecting the data, reviewing outputs regularly, using security-oriented tools and ensuring that no corporate classified and sensitive data are contained in datasets. The next category of risks regards the model's vulnerabilities that malicious actors can exploit.

### *Adversarial Machine Learning*

AI has specific vulnerabilities compared to traditional IT systems.<sup>234</sup> The most common AI vulnerabilities and recommendations to enhance security for deployed AI models will be described below. Each time the CAF Cyber Commander deploys an AI model, the organization's attack surface<sup>235</sup> will increase.<sup>236</sup> Malicious actors target AI model vulnerabilities using Adversarial Machine Learning (AML) techniques, which exploit vulnerabilities in ML components, such as software, hardware, workflows, and supply chains.<sup>237</sup> AML techniques intend to deceive or force the AI model to return unintended information.<sup>238</sup> They also aim to influence output with the feed of specific input.<sup>239</sup> Gartner<sup>240</sup> published its first Adversarial Machine Learning (AML) report in February 2019.<sup>241</sup> The most common AML techniques discussed below are adversarial inputs, poisoning training data, and model extraction attacks.<sup>242</sup>

---

<sup>232</sup> *Ibid.*

<sup>233</sup> "Data Prep - Construct Your Dataset."

<sup>234</sup> Communications Security Establishment, *Canadian Centre for Cyber Security - National Cyber Threat Assessment 2023-2024* (His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2022), 20.

<sup>235</sup> Attack surface is defined as "the sum of vulnerabilities, pathways, or methods—sometimes called attack vectors—that hackers can use to gain unauthorized access to the network or sensitive data, or to carry out a cyberattack." Source: "What is an Attack Surface?" last modified n.d., accessed March 14, 2024, <https://www.ibm.com/topics/attack-surface>.

<sup>236</sup> "Mitre Atlas™," accessed Dec 5, 2023, <https://atlas.mitre.org/>.

<sup>237</sup> UK National Cyber Security Centre (GCHQ), *Guidelines for Secure AI System Development* (Washington, D.C.: UK NCSC, US Cybersecurity and Infrastructure Security Agency (CISA), 2023), 20.

<sup>238</sup> Communications Security Establishment, *Canadian Centre for Cyber Security - National Cyber Threat Assessment 2023-2024*, 20.

<sup>239</sup> AI-Dosari, "Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges," 5.

<sup>240</sup> Gartner is leading industry market research firm. Source: "Life at Gartner," last modified n.d., accessed May 4, 2024, <https://www.gartner.com/en>.

<sup>241</sup> Ram Shankar Siva Kumar et al., "Adversarial Machine Learning-Industry Perspectives," 978-1-7281-9346-5 (-05, 2020), , 69. doi:10.1109/spw50608.2020.00028. <https://ieeexplore-ieee-org.cfc.idm.oclc.org/document/9283867>.

<sup>242</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 16.

### *Adversarial Inputs*

Malicious actors will modify the input to force the AI model to make a wrong output prediction to evade detection.<sup>243</sup> A common technique used is the Malicious Generative Adversarial Networks (MalGAN). It creates adversarial inputs that bypass the AI-enabled detection models for instance.<sup>244</sup>

### *Poisoning Training Datasets*

The training datasets are another component of the AI model that is vulnerable to malicious actors. Malicious actors will try to target the AI model's training phase by modifying the training datasets to influence its output.<sup>245</sup> False data will be injected into the training dataset before the training phase so that the AI model will produce incorrect output.<sup>246</sup>

### *Model Extraction Attacks*

The model extraction attack intends to recover the training datasets via black-box examination.<sup>247</sup> The malicious actors will reproduce the training sets.<sup>248</sup> Malicious actors can, with reversing techniques, learn how ML algorithms work and know what the AI model is explicitly looking for.<sup>249</sup> With this knowledge, malicious actors can deceive the AI models.

### *Mitigation Measures Against AML*

This section recommends measures to harden the defence of AI models. Defending AI-enabled systems against malicious actors will become vital for organizations that leverage them for their operations. Otherwise, they will not be able to rely on them, and this could cause significant damage to the organization. Meanwhile, it is necessary to know that AI security is not as developed as traditional cybersecurity.

### *Gap Between Traditional Cyber Security and AI-enabled Systems Security Practices*

There is a gap between traditional cyber security and AI-enabled systems security practices. In 2020, the practice of tracking and scoring ML vulnerabilities using the Common Vulnerability Scoring System was poorly developed.<sup>250</sup> Moreover, there is a gap between methods used to defend traditional IT systems and AI models. To highlight the gap between traditional cyber security and security for AI models, the organization that developed the MITRE

---

<sup>243</sup> *Ibid.*

<sup>244</sup> *Ibid.*

<sup>245</sup> *Ibid.*

<sup>246</sup> Canadian Centre for Cyber Security, *Artificial Intelligence - ITSAP.00.040*.

<sup>247</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 16.

<sup>248</sup> Canadian Centre for Cyber Security, *Artificial Intelligence - ITSAP.00.040*.

<sup>249</sup> Truong, "Artificial Intelligence in the Cyber Domain: Offense and Defense," 16.

<sup>250</sup> Siva Kumar, "Adversarial Machine Learning-Industry Perspectives," 73.

ATT&CK® Framework used for cybersecurity purposes had to develop a distinct framework for AI called MITRE ATLAS™.

### *Threat Modelling & MITRE ATLAS™*

The MITRE ATT&CK® Framework provided a knowledge base of adversary tactics, techniques, and procedures based on real-world events. Given the growing number of vulnerabilities in AI models and the expansion of organizations' attack surface with the integration of AI-enabled systems, *The MITRE Corporation* is proposing a similar framework for AI-enabled systems, the MITRE ATLAS™ (Adversarial Threat Landscape for Artificial-Intelligence Systems).<sup>251</sup> MITRE ATLAS™ is, like the MITRE ATT&CK® Framework, a knowledge base of adversary tactics and techniques from real-world attack observations or achievements from AI red teams or security researchers.<sup>252</sup> Before deploying an AI model, it is recommended that a threat modelling be established. The MITRE ATLAS™ is an excellent tool for accomplishing this task. Furthermore, it would be necessary to conduct red teaming on the AI model.

### *AI Red Teaming*

Red teaming, commonly called penetration tests, is a current practice in traditional cybersecurity in which a team deliberately tries to exploit a system to identify its vulnerabilities and evaluate the risks of being exploited.<sup>253</sup> Using the red teaming approach is an efficient way to reduce the risk when deploying AI-enabled systems.<sup>254</sup> Firstly, an AI red team should assess the output risks, especially if a Gen AI model will be deployed. To assess the output risks, the red team will attempt to get the model to generate harmful content.<sup>255</sup> Secondly, red teams must use AML techniques against the model in addition to assessing the output risk. As mentioned in the previous section, red teams could use tools such as MITRE ATLAS™ to model the threat. Currently, organizations in the industry have either in-house or external red teams. For instance, Google has its own internal red teams, whereas OpenAI is leveraging the concept of external red teams.<sup>256</sup> Either way, before deploying an AI-enabled system, assessing the risk with a red teaming approach is recommended.

---

<sup>251</sup> "Mitre Atlas™."

<sup>252</sup> *Ibid.*

<sup>253</sup> Siva Kumar, "Adversarial Machine Learning-Industry Perspectives," 73.

<sup>254</sup> "How to Red Team a Gen AI Model," last modified January 4, accessed March 14, 2024, <https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model>.

<sup>255</sup> *Ibid.*

<sup>256</sup> *Ibid.*

### *Secure AI Model Lifecycle*

The UK National Cyber Security Centre (GCHQ) recommends that the security of the AI model be a core requirement during the development phase and throughout the life cycle.<sup>257</sup> This is why the last step of the proposed 10-step framework described at the beginning of this chapter is dedicated to defending the AI model. The Guidelines for Secure AI System Development from the UK National Cyber Security Centre (GCHQ) propose securing four critical areas during the AI system life cycle: the design, development, deployment, and secure operation and maintenance.<sup>258</sup>

Furthermore, building AI models requires secure coding practices to reduce exploitable vulnerabilities, as in writing traditional software.<sup>259</sup> The most common toolkits for AI models are Pytorch, Keras, and TensorFlow, which include best practices for secure coding.<sup>260</sup> TensorFlow also provides tools for testing the model against AML.<sup>261</sup>

In-house model development could potentially reduce the risk of common AML, which could be facilitated by COTS solutions.<sup>262</sup> Developing an AI model in-house will not eliminate vulnerabilities; however, it might reduce forms of attack requiring internet connections or cloud system breaches to access the model's training dataset.<sup>263</sup>

### *DND/CAF AI Centre (DCAIC)*

*The DND and CAF AI Strategy* proposes creating a centre of excellence for developing AI called the DND/CAF AI Centre (DCAIC). DCAIC would be a hub of AI expertise and provide a secure environment for AI development, experimentation, and testing at all classification levels.<sup>264</sup> Establishing the DCAIC by the DND and CAF would improve and facilitate the defence of AI-enabled systems and AI models. Moreover, it would also be responsible for labelled datasets, which might reduce AML vulnerabilities.<sup>265</sup> Therefore, it is recommended to coordinate with the Digital Transformation Officer (DTO) and, if it is created, with the DCAIC before initiating any AI-enabled project.

To summarize, the defence of AI models requires a different approach from traditional IT security systems. It is essential to be cognizant that each newly deployed AI-enabled system will increase the DND and the CAF's attack surface. Here are a few suggestions to harden the

---

<sup>257</sup> UK National Cyber Security Centre (GCHQ), *Guidelines for Secure AI System Development*, 5.

<sup>258</sup> *Ibid.*

<sup>259</sup> Siva Kumar, "Adversarial Machine Learning-Industry Perspectives," 71.

<sup>260</sup> *Ibid.*

<sup>261</sup> *Ibid.*

<sup>262</sup> Taddeo, "Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword," 560.

<sup>263</sup> *Ibid.*

<sup>264</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 15.

<sup>265</sup> *Ibid.*

defence of an AI model. Firstly, securing the whole AI model lifecycle from its design to its end-of-life is recommended. Secondly, it is suggested that threat modelling be established using the tool MITRE ATLAS™. Thirdly, red teaming will help identify output risks and AML vulnerabilities in the model. Finally, reach out to the DTO and, if it exists, the DCAIC. In addition to defending its AI models, the CAF Cyber Command will have to follow regulations and compliance, which will be detailed in the next section.

### Section 3.3 – Regulations and Compliance

Before creating any AI models, it is essential to be aware of the Government of Canada's regulations and compliance regarding AI. This section will discuss the requirements extracted from *The DND and CAF AI Strategy*, the *Directive on Automated Decision-Making*, the *Artificial Intelligence and Data Act*, the *Pan-Canadian Artificial Intelligence Strategy*, and the *Guide on the use of Generative AI*.

#### The DND and CAF AI Strategy Requirements

Chapter 1 presented *The DND and CAF AI Strategy*, which was published in March 2024 and aimed to establish a vision for the DND and CAF's adoption of AI. The strategy also proposes five LoEs to achieve this vision.<sup>266</sup> Following all LoEs will ensure that all DND and CAF's requirements regarding AI are met. Here are some examples of requirements that must be satisfied.

#### Legal, Inclusive, Ethical, and Safe

Any AI models developed or procured by the DND and CAF must be legal, inclusive, ethical, and safe.<sup>267</sup> These requirements must be embedded within each stage of the AI model lifecycle: “design, development, validation and certification, procurement, and deployment of AI systems to their eventual decommissioning.”<sup>268</sup> DND and CAD organizations wanting to leverage AI must develop guidance to ensure AI's responsible and transparent employment, especially when laws and policies are yet to be developed.<sup>269</sup>

*The DND and CAF AI Strategy* emphasizes, under its LoE 3: Ethics, Safety, and Trust, that all new AI-enabled technologies developed, procured, or implemented must be in accordance with laws, policies, and guidelines.<sup>270</sup> Table 1 below enumerates the most common laws, policies, guidelines, and guidance that DND and CAF must comply with to leverage AI.

---

<sup>266</sup> *Ibid.*, 32.

<sup>267</sup> *Ibid.*, 22.

<sup>268</sup> *Ibid.*, 18.

<sup>269</sup> *Ibid.*, 11.

<sup>270</sup> *Ibid.*, 19.

**Table 3.3 – Laws, Policies, Guidelines, Guidance, and Commitment for the Development and Implementation of AI-enabled Technologies**

Regulation Type	Name
Laws	Canadian Law
	International Law
Regulations and Government of Canada Policies	Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems
	Directive on Automated Decision-Making
Guidelines	Guide on the use of Generative AI
Guidance from Governmental Review Entities	National Security and Intelligence Committee of Parliamentarians (NSICOP)
	National Security and Intelligence Review Agency (NSIRA)
	Office of the Privacy Commissioner (OPC)
Commitment	Gender-Based Analysis Plus (GBA+)

Source: National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 19.

Emphasis must be applied to each AI project's ethical and security concerns. It is required to identify sources of unintended harm and vulnerabilities in the AI model and mitigation measures to address these shortcomings.<sup>271</sup> Moreover, all AI projects must incorporate GBA+ to address the needs of different groups and avoid harm generated by biased data or algorithms.<sup>272</sup> Finally, clear terms of reference must be established, which define personnel's roles, responsibilities, and authorities in developing or procuring AI.<sup>273</sup>

#### Interoperability and Partnerships

AI-enabled systems must be adaptable and interoperable with DND and CAF partners such as the Technical Cooperation Program (TTCP) and allies such as the Five Eyes and NATO.<sup>274</sup> Furthermore, CAF organizations aspiring to leverage AI must be prepared to work with industry, academia, government departments, and non-traditional partners.<sup>275</sup> If CAF Cyber Command leadership wishes to develop AI models, it should consult key programs, such as the DRDC research centres and the MINDS and the IDEaS programs.<sup>276</sup> Consequently, it is

<sup>271</sup> *Ibid.*, 17.

<sup>272</sup> *Ibid.*, 18.

<sup>273</sup> *Ibid.*

<sup>274</sup> *Ibid.*, 32.

<sup>275</sup> *Ibid.*

<sup>276</sup> *Ibid.*, 25.

recommended that one be familiar with *The DND and CAF AI Strategy* before leveraging AI models. Any AI project should be aligned with its five LoEs.

### Directive on Automated Decision-Making

The *Directive on Automated Decision-Making* took effect on April 1, 2019, and required compliance no later than April 1, 2020.<sup>277</sup> The Treasury Board of Canada Secretariat is responsible for this directive, which applies to all institutions subject to the *Policy on Service and Digital*. This directive aims to ensure that automated decision systems are deployed with efficiency and minimum risks to clients, federal organizations, and Canadian society and in accordance with Canadian law.<sup>278</sup> This directive must be respected by federal organizations that intend to leverage AI technologies and automated decision systems, such as rules-based systems, regression, machine learning and deep learning.<sup>279</sup>

This directive established the following requirements for automated decision systems, which are fully described on their website: Algorithmic Impact Assessment, Transparency, Quality Assurance, Recourse, and Reporting.<sup>280</sup> There are consequences for not complying with the directive allowed by the *Financial Administration Act*.<sup>281</sup>

### Algorithm Impact Assessment

The Algorithm Impact Assessment (AIA) is a mandatory risk assessment tool maintained by the Office of the Chief Information Officer (OCIO) at the Treasury Board of Canada Secretariat (TBS).<sup>282</sup> As mentioned at the beginning of this chapter, this requirement should be completed during Step 6 – Validating the Model from the 10-step framework. AIA is a questionnaire that establishes the impact level of an automated decision system with 51 risk and 34 mitigation questions.<sup>283</sup> The AIA provides a score after evaluating several factors, such as the data, system design, decision type, algorithms, and impact.<sup>284</sup> Four levels of impact establish the mitigation measures required by the *Directive on Automated Decision-Making* to be implemented before deploying the system.<sup>285</sup> The final results from the AIA must be released in

---

<sup>277</sup> "Directive on Automated Decision-Making," last modified April 25, accessed January 12, 2024, <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

<sup>278</sup> *Ibid.*

<sup>279</sup> Definition of Automated Decision System: "Any technology that either assists or replaces the judgment of human decision-makers. These systems draw from fields like statistics, linguistics and computer science, and use techniques such as rules-based systems, regression, predictive analytics, machine learning, deep learning, and neural nets." Source: "Directive on Automated Decision-Making," last modified April 25, accessed January 12, 2024, <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.

<sup>280</sup> "Directive on Automated Decision-Making."

<sup>281</sup> *Ibid.*

<sup>282</sup> "Algorithmic Impact Assessment Tool."

<sup>283</sup> *Ibid.*

<sup>284</sup> *Ibid.*

<sup>285</sup> *Ibid.*

both languages on the Open Government Portal.<sup>286</sup> Further to completing the AIA, consulting with legal services must be done at the beginning of the project.<sup>287</sup> Consequently, the CAF Cyber Command leadership and staff must ensure that all AI projects have completed an AIA and comply with the *Directive on Automated Decision-Making* requirements. In addition to this directive, if Gen AI models are leveraged, the *Guide on the use of Generative AI* should be consulted.

### Guide on the use of Generative AI

This guidance, along with existing federal laws and policies, is an additional tool for government institutions that aspire to leverage Gen AI specifically.<sup>288</sup> This guide recommends that Federal institutions leverage Gen AI to help and support their employees, not to replace them.<sup>289</sup> Federal institutions should ensure that Gen AI tools align with FASTER principles, which are “Fair, Accountable, Secure, Transparent, Educated, and Relevant,”<sup>290</sup> and comply with laws and policies.<sup>291</sup> Important to note the guide refers to the *Directive on Automated Decision-Making*, which is also applied to Gen AI technologies.<sup>292</sup>

### Documentation Requirements

Likewise, the guide discusses documentation requirements for federal institutions that want to leverage Gen AI. It refers to the *Directive on Service and Digital*, which requires that Government of Canada employees document their activities that create business value.<sup>293</sup> If the CAF Cyber Command leadership wants to leverage Gen-AI, it must provide records of decisions about the development and deployment of the tool.<sup>294</sup> All steps taken to ensure appropriate and accurate output from the tool must also be recorded.<sup>295</sup> The guide provides information about “potential issues and best practices,” such as “Protection of information, Bias, Quality, Public servant autonomy, Legal risks, Distinguishing humans from machines, and Environmental

---

<sup>286</sup> *Ibid.*

<sup>287</sup> *Ibid.*

<sup>288</sup> "Guide on the use of Generative AI," last modified February 20, accessed March 21, 2024, <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html>.

<sup>289</sup> "Guide on the use of Generative AI," last modified February 20, accessed March 21, 2024, <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html>.

<sup>290</sup> *Ibid.*

<sup>291</sup> *Ibid.*

<sup>292</sup> *Ibid.*

<sup>293</sup> "Directive on Service and Digital," last modified January 10, accessed March 17, 2024, <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32601>.

<sup>294</sup> "Guide on the use of Generative AI."

<sup>295</sup> *Ibid.*

impacts.”<sup>296</sup> In summary, the *Guide on the use of Generative AI* is an additional guideline that the CAF Cyber Command should comply with if Gen AI is leveraged.

#### Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems

Besides the compliances and regulations, the CAF Cyber Command leadership and staff should also follow the *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*. It proposes managers of Gen AI models commit to the following principles and their specific measures: “Accountability, Safety, Fairness and Equity, Transparency, Human Oversight and Monitoring, and Validity and Robustness.”<sup>297</sup> A detailed checklist for each principle is available on the website. Signatories of the code of conduct, such as CGI, BlackBerry, IBM or TELUS, commit to supporting a robust and responsible AI ecosystem in Canada.<sup>298</sup>

#### Artificial Intelligence and Data Act

In June 2022, the *Artificial Intelligence and Data Act* (AIDA) was tabled as the *Digital Charter Implementation Act*, part of Bill C-27, to protect Canadians and guarantee the responsible development of AI.<sup>299</sup> AIDA is aligned with international norms, such as the EU AI Act, the Organization of Economic Co-operation and Development AI Principles and the US National Institute of Standard and Technology (NIST).<sup>300</sup> The Minister of Innovation, Science, and Industry is authorized to administer and enforce the AIDA.<sup>301</sup> AIDA was one of the first Canadian regulatory frameworks for AI.<sup>302</sup> AIDA is reinforced with three types of consequences for non-compliance: administrative monetary penalties (AMPs), prosecution of regulatory offences, and actual criminal offences.<sup>303</sup> The AIDA created three new criminal offences, separated from the regulatory obligations.<sup>304</sup> AIDA is primarily concerned with adopting responsible AI technologies and developing positive innovations by Canadian businesses and organizations.<sup>305</sup> Thus, the CAF Cyber Command leadership and staff should be familiar with the AIDA and follow its principles to comply with the law.

---

<sup>296</sup> *Ibid.*

<sup>297</sup> "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems," last modified January 12, accessed March 17, 2024, <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.

<sup>298</sup> *Ibid.*

<sup>299</sup> "The Artificial Intelligence and Data Act (AIDA) – Companion Document," last modified March 13, accessed March 16, 2024, <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>.

<sup>300</sup> *Ibid.*

<sup>301</sup> *Ibid.*

<sup>302</sup> *Ibid.*

<sup>303</sup> *Ibid.*

<sup>304</sup> *Ibid.*

<sup>305</sup> *Ibid.*

## Pan-Canadian Artificial Intelligence Strategy

The *Pan-Canadian Artificial Intelligence Strategy* is not a regulation or compliance; however, it is an excellent source of information for which the Government of Canada is investing financially for those who intend to leverage AI technologies. The Government of Canada partnered with Canadian organizations, such as CIFAR, AMII, Mila, Vector Institute, Global Innovation Clusters, SCC / CCN, and Digital Research Alliance of Canada, to implement the *Pan-Canadian Artificial Strategy*.<sup>306</sup> This strategy aims to support the adoption of AI across Canada and society.<sup>307</sup> It includes three pillars. The first pillar is commercialization, which provides for three National Artificial Intelligence Institutes that facilitate the translation of research in AI into commercial applications: Amii in Edmonton, Mila in Montreal, and the Vector Institute in Toronto.<sup>308</sup> The second pillar concerns efforts to develop and adopt standards for AI technologies.<sup>309</sup> The third pillar intends to improve talent and research through the Canadian Institute for Advanced Research (CIFAR) and Digital Research Alliance.<sup>310</sup> This is relevant for the CAF Cyber Command leadership and staff to know about these institutions that could support developing and implementing AI models.

### Summary

This section identified Government of Canada regulations and compliance regarding AI technologies. All AI models developed and implemented by the DND and CAF will be required to comply with the *Directive on Automated Decision-Making*, *Algorithmic Impact Assessment*, *Guide on the use of Generative AI*, *Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems*, and the *Artificial Intelligence and Data Act*. The Government of Canada website called *Responsible for Artificial Intelligence (AI)* provides good information and services about AI.<sup>311</sup> Before initiating any AI project, CAF Cyber Command leadership and staff should study requirements from *The DND and CAF AI Strategy* and then move on to the key regulations.

### Section 3.4 – XAI

This section is about an emerging research field that aims to improve transparency and interpretability issues in AI. Since AI-enabled systems are more often integrated into decision-making processes, the interpretability and explainability of their output are becoming critical.

---

<sup>306</sup> "Pan-Canadian Artificial Intelligence Strategy," last modified July 20, accessed March 16, 2024, <https://ised-isde.canada.ca/site/ai-strategy/en>.

<sup>307</sup> *Ibid.*

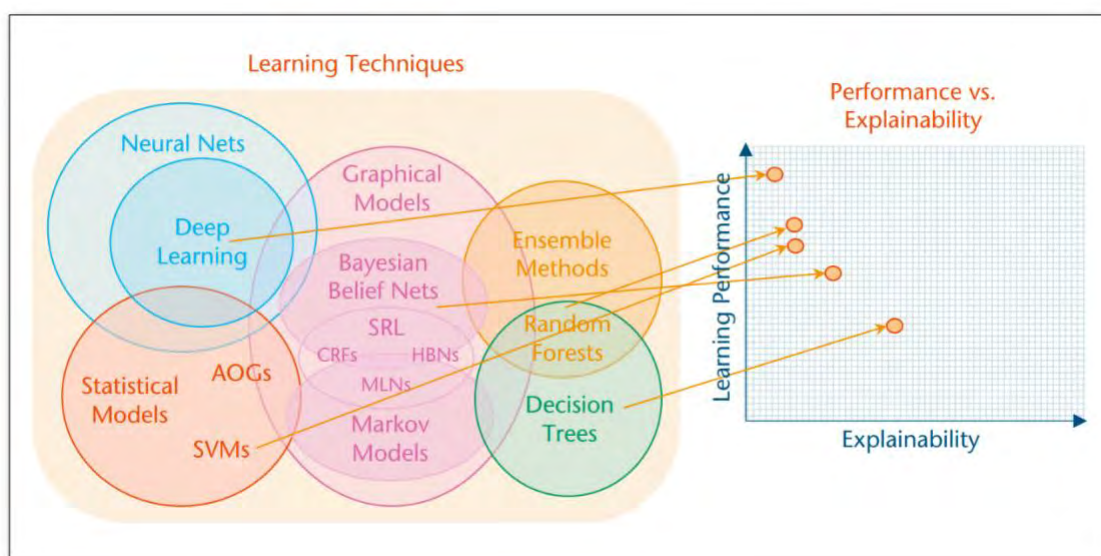
<sup>308</sup> *Ibid.*

<sup>309</sup> *Ibid.*

<sup>310</sup> *Ibid.*

<sup>311</sup> "Responsible use of Artificial Intelligence (AI)," last modified February 2, accessed March 17, 202, <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>.

Interpretability refers to “the degree to which a machine learning model can accurately link a cause (input) with an outcome (output).”<sup>312</sup> On the other hand, explainability refers “to the degree to which the internal working of a machine learning or deep learning system can be articulated in human words.”<sup>313</sup> The need to understand and explain the output predictions is increasing.<sup>314</sup> The current problem in AI is the compromise between model performance and model transparency.<sup>315</sup> For example, DL models’ output has a high level of accuracy, but they are the least explainable. On the other hand, simpler ML, such as the decision tree model, is the least accurate but the most explainable.<sup>316</sup> Figure 3.6 provides examples of learning techniques and their explainability versus learning performance.



**Figure 3.6 – Performance versus Explainability**

Source: Gunning, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine* 40, no. 2 (2019), 46.

In the literature, the least explainable and transparent models are frequently referred to as “black box” models.<sup>317</sup> These models could have even modified or adjusted some parameters automatically during their learning process without human intervention, which makes their

<sup>312</sup> "Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review," last modified February 28, accessed March 23, 2024, <https://www.statcan.gc.ca/en/data-science/network/explainable-learning>.

<sup>313</sup> *Ibid.*

<sup>314</sup> Ahmed, "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence," 2.

<sup>315</sup> Rudresh Dwivedi et al., "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," *ACM Computing Surveys* 55, no. 9 (January, 2023), 31. doi:10.1145/3561048. <https://dl.acm.org/doi/10.1145/3561048>.

<sup>316</sup> David Gunning and David W. Aha, "DARPA's Explainable Artificial Intelligence Program," *AI Magazine* 40, no. 2 (2019), 45. <https://login.cfc.idm.oclc.org/login?url=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fdarpas-explainable-artificial-intelligence%2Fdocview%2F2258093718%2Fse-2%3Faccountid%3D9867>.

<sup>317</sup> Zhang, "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities," 1042.

explainability even more difficult.<sup>318</sup> To counter these problems, several organizations have been working on potential solutions that will be discussed below.

## DARPA XAI

In 2015, the Defense Advanced Research Projects Agency (DARPA) initiated the development of a new program called “Explainable Artificial Intelligence,” also called XAI.<sup>319</sup> Other organizations also started to develop XAI simultaneously; however, DARPA centralized the topics and developed a single system for evaluating explanations.<sup>320</sup> DARPA defines XAI as:

AI systems that can explain their rationale to a human user, characterize their strengths and weaknesses, and convey an understanding of how they will behave in the future.”<sup>321</sup> Many motivations behind the emergence of XAI are to improve trust, causality, transferability and informativeness toward the algorithm.<sup>322</sup>

The DARPA’s XIA program aimed to create new or modified ML models for the end user who depends on an AI model's decisions and recommendations. Therefore, the user needs to understand the model’s strengths and weaknesses, how it provided a particular output, and how it will behave in the future.<sup>323</sup>

## NIST 4 XAI Principles

XAI is a multidisciplinary field that involves computer science, engineering, and psychology.<sup>324</sup> Figure 3.8 describes the four XAI principles proposed by the National Institute of Standards and Technology (NIST) of the U.S. Department of Commerce: explanation, meaningfulness, explanation accuracy, and knowledge limits.<sup>325</sup>

---

<sup>318</sup> *Ibid.*

<sup>319</sup> Mayuri Mehta, Vasile Palade and Indranath Chatterjee, "Explainable AI - Foundations, Methodologies and Applications," *Springer* 232 (2023), 26. doi:10.1007/978-3-031-12807-3. <https://doi.org/10.1007/978-3-031-12807-3>.

<sup>320</sup> *Ibid.*

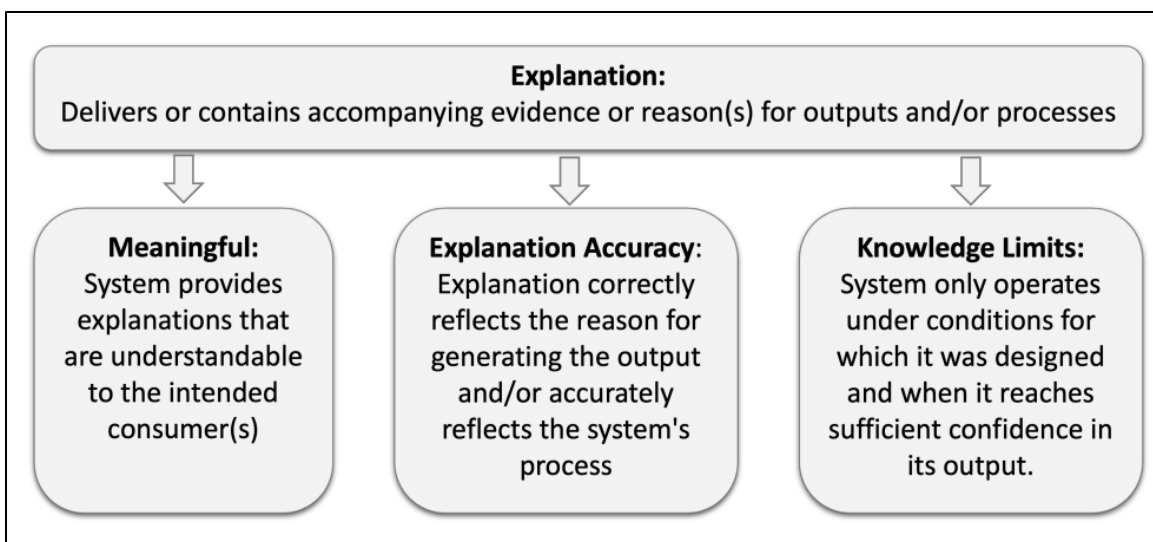
<sup>321</sup> Gunning, "DARPA's Explainable Artificial Intelligence Program," 45.

<sup>322</sup> Mehta, "Explainable AI - Foundations, Methodologies and Applications," 26.

<sup>323</sup> Gunning, "DARPA's Explainable Artificial Intelligence Program," 45.

<sup>324</sup> P. Jonathon Phillips et al., "Four Principles of Explainable Artificial Intelligence," *NIST Interagency/Internal Report (NISTIR)* 8312 (September 29, 2021), i. doi:10.6028/nist.ir.8312. <https://doi.org/10.6028/NIST.IR.8312>.

<sup>325</sup> *Ibid.*, 3.



**Figure 3.7 – NIST XAI 4 Principles**

Source: Phillips, "Four Principles of Explainable Artificial Intelligence," 3.

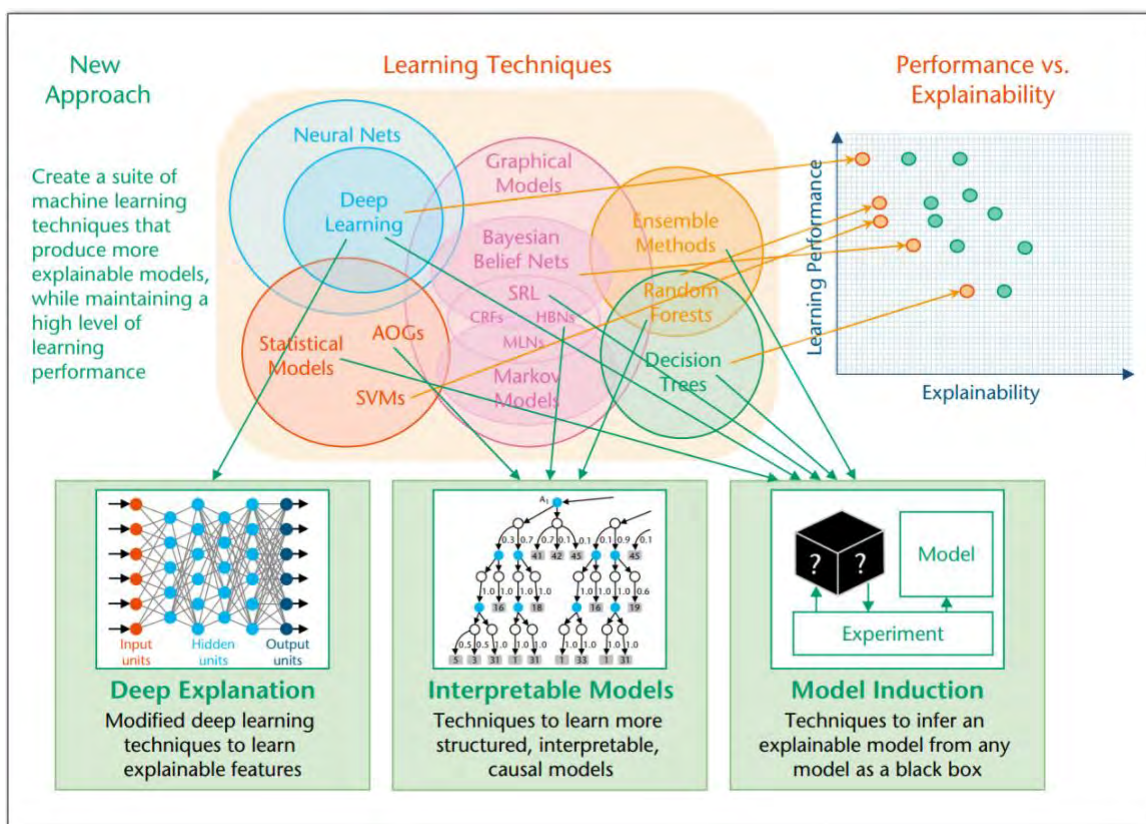
NIST proposed these principles to serve as guidance for assessing the explanation required by the user. To improve a model's explainability, it is recommended to meet as many of these four principles as possible.<sup>326</sup>

#### XAI Techniques

The scope of this study will not cover all XAI techniques in detail; it will cover a few broad concepts. In 2017, DARPA proposed three strategies to improve explainability: deep explanation, interpretable models, and model induction, as described in Figure 3.9.<sup>327</sup>

<sup>326</sup> *Ibid.*, 5.

<sup>327</sup> Gunning, "DARPA's Explainable Artificial Intelligence Program," 45.

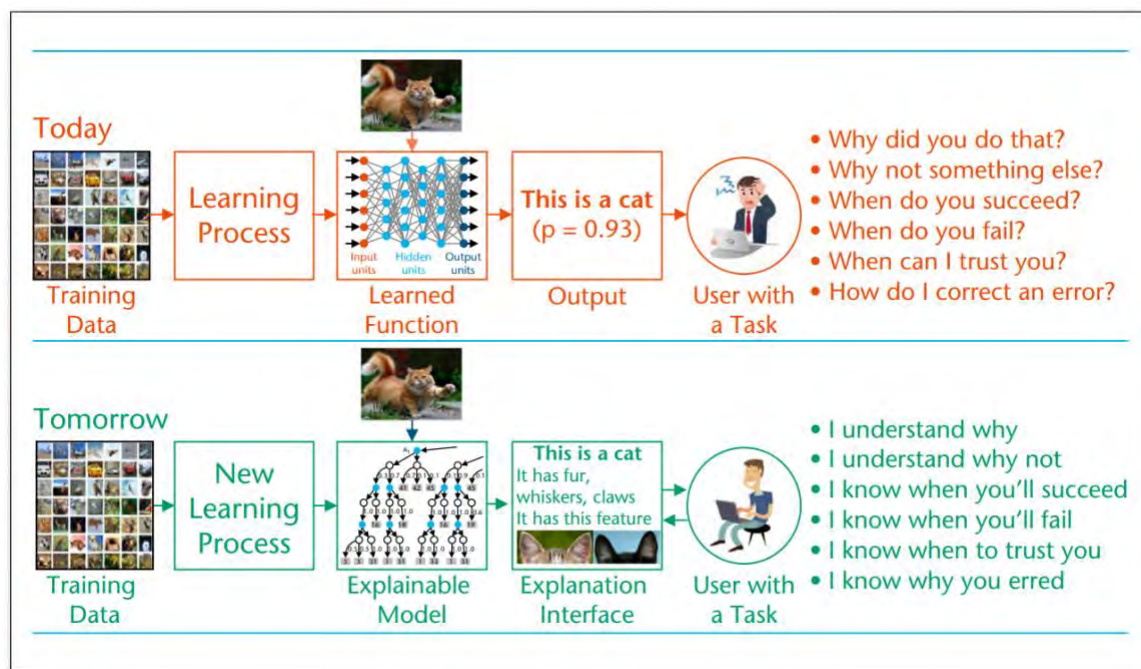


**Figure 3.8 – DARPA's 3 Strategies to Improve Explainability**

Source: Gunning, "DARPA's Explainable Artificial Intelligence Program," 47.

Figure 3.9 was a model proposed by DARPA. It illustrates the difference between a current ML model and what XAI's new ML model would look like. Some key differences are that new learning processes replace the learning algorithm, and an explainable interface provides additional information on the output for the user.<sup>328</sup>

<sup>328</sup> *Ibid.*, 48.



**Figure 3.9 – DARPA XAI Concept**

Source: Gunning, "DARPA's Explainable Artificial Intelligence Program," 68.

Several algorithms have been developed to explain AI models, which can be divided into two categories: self-interpretable models and post-hoc explanations.<sup>329</sup>

#### *Self-Interpretable Models or White Box Models*

The Self-Interpretable category includes models that are by themselves an explanation.<sup>330</sup> These models can give a global explanation for the entire model and a local explanation for each decision made by the model.<sup>331</sup> Common self-interpretable models are shallow ML using decision trees and regression algorithms.<sup>332</sup> As mentioned previously and pointed out in Figure 4, these types of models have a high level of explainability but a low level of accuracy. This category is also called white box since its mechanisms are transparent, contrary to a black box model.<sup>333</sup>

<sup>329</sup> Phillips, "Four Principles of Explainable Artificial Intelligence," 12.

<sup>330</sup> *Ibid.*

<sup>331</sup> *Ibid.*

<sup>332</sup> *Ibid.*

<sup>333</sup> Dwivedi, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," 6.

## Post-Hoc Explanations

The Post-hoc Explanations category is divided into two sub-categories: local explanation and global explanation.<sup>334</sup> Global explanation aims to assess the entire model, not just a specific decision. It includes the model's general behaviour, variables, dependency, and their interactions.<sup>335</sup> The most common techniques for global explanation are the feature importance and the partial dependence plots (PDP).<sup>336</sup> Local explanations will provide explanations for individual output predictions.<sup>337</sup> Local explanations will be done on a subregion of a particular data point.<sup>338</sup> Local Interpretable Model Agnostic Explanation (LIME) and the Shapley Additive exPlanation (SHAP) value are two common techniques for conducting local explanations.<sup>339</sup> LIME will select a decision that requires an explanation. It will build an interpretable model, which is usually a logistic regression, for the local decision by assessing nearby data points.<sup>340</sup> SHAP uses the game theory approach<sup>341</sup> to explain the output.<sup>342</sup> The SHAP can explain the output by computing each feature to the prediction.<sup>343</sup> Moreover, the SHAP can provide local and global explanations but requires more computational resources than the LIME techniques.<sup>344</sup>

## Risks and Challenges Related to XIA

As mentioned previously, XIA aims to provide explanations so humans can understand how the model obtains the output.<sup>345</sup> The article *Viewpoint Explainable AI – Opening the black box or Pandora's Box?* highlights some risks related to using XAI.<sup>346</sup> As the article mentioned, opening an AI black box model could be the equivalent of opening a Pandora Box.<sup>347</sup> The common risks of using XAI are revealing flaws in datasets and model designs, classified and sensitive information, organizational processes and even intellectual property.<sup>348</sup> Even if XAI cannot fully open a black box AI model nowadays, it might become the case in years to come.

<sup>334</sup> Phillips, "Four Principles of Explainable Artificial Intelligence," 12.

<sup>335</sup> Dwivedi, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," 8.

<sup>336</sup> *Ibid.*

<sup>337</sup> *Ibid.*

<sup>338</sup> *Ibid.*

<sup>339</sup> *Ibid.*

<sup>340</sup> Phillips, "Four Principles of Explainable Artificial Intelligence," 13.

<sup>341</sup> Game Theory: "is a theoretical framework for social interactions with competing actors. It's the study of optimum decision-making by independent and competing agents in a strategic context. A "game" is any scenario in which there are many decisions makers, each of whom seeks to maximize their outcomes. The optimal choice will be influenced by the decisions of others. The game determines the participants' identities, preferences, and possible tactics, as well as how these strategies influence the result." Source: "Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review."

<sup>342</sup> Dwivedi, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," 20.

<sup>343</sup> "Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review."

<sup>344</sup> Dwivedi, "Explainable AI (XAI): Core Ideas, Techniques, and Solutions," 21.

<sup>345</sup> Storey, "Viewpoint - Explainable AI," 27.

<sup>346</sup> *Ibid.*

<sup>347</sup> A Pandora Box means that "In Greek mythology, Pandora's Box refers to a container of evils that are unleashed and cannot be contained once the box is opened." Source: Storey, "Viewpoint - Explainable AI," 28.

<sup>348</sup> *Ibid.*

When developing and integrating AI models, organizations should remember that they could be opened and reveal sensitive information about their processes, data, classified and sensitive information, and intellectual property.<sup>349</sup> When building any AI model, being cognizant of these risks today is crucial.

Consequently, CAF Cyber Command leadership and staff should consider using XAI techniques to mitigate output risks, as discussed in Section 3.2. Eventually, more AI models will be integrated into decision-making processes. It is assumed that their output will need to be interpretable and explainable. XAI may be a solution to these problems.

### **Chapter 3 Summary**

This chapter defined the CAF Cyber Command leadership and staff's responsibilities regarding leveraging AI for cyber operations. It was divided into four sections. The first section proposed a 10-step framework for creating in-house AI models. The second section concerned the defence needed for AI models against their output risks and AML. The third section discussed regulations and compliance to be followed when developing and implementing an AI-enabled system. The last section presented XAI, which is an emergent field to solve AI transparency and trust issues. The next chapter will explain how the CAF Cyber Command can leverage AI for Defensive Cyber Operations.

---

<sup>349</sup> *Ibid.*

## CHAPTER 4 – AI FOR DEFENSIVE CYBER OPERATIONS

This chapter will explain how the CAF Cyber Command can leverage AI to conduct DCO, and it will be divided into four sections. The first section will define DCO, its requirements and key activities. The second section will highlight key considerations about leveraging AI for DCO. The third and fourth sections will propose solutions to leverage AI for the planning and executing DCO, respectively. The scope of this chapter is not to provide detailed AI solutions for DCO but key considerations and a few examples of applications where AI could be leveraged. Besides, some proposed solutions in this chapter might become outdated rapidly due to the speed of development of the cyber domain and AI fields.

### Section 4.1 – Defensive Cyber Operations

#### DCO Definitions

Nowadays, CAF and its allies and adversaries have developed full-spectrum cyber operations capabilities.<sup>350</sup> The *CAF JDN 2017-02 - Cyber Operations*, published in 2017, is the only available CAF cyber domain doctrine. This doctrine describes the following types of cyber operations: defensive cyber operations (DCO), offensive cyber operations (OCO), and support cyber operations.<sup>351</sup> DCOs distinguish themselves from traditional cyber security by being threat—and mission-oriented. They are threat-oriented since they are planned and executed in response to a specific threat that has compromised or threatened to compromise CAF-owned systems. They are also mission-oriented because they prioritize key terrain and defend military operations against adversary activities in cyberspace.<sup>352</sup> The *CAF JDN 2017-02 – Cyber Operations* defines DCOs as: “A defensive operation conducted in or through cyberspace to detect, defeat and/or mitigate offensive and exploitive actions to maintain freedom of action. A DCO may include internal defensive measures<sup>353</sup> and response action.”<sup>354</sup> As mentioned in the introduction, this study will focus solely on DCO-IDM and not discuss DCO-RA. The next subsection will detail the key requirements and activities for conducting DCOs.

---

<sup>350</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-4.

<sup>351</sup> *Ibid.*, 4-2.

<sup>352</sup> *Ibid.*, 4-6.

<sup>353</sup> Defensive Cyber Operations – Internal defensive measures (DCO-IDM) are defined as: “In DCOs, measures and activities conducted within one’s own cyberspace to ensure freedom of action.” whereas Defensive Cyber Operations – Responsive Actions (DCO-RA) are defined as “In DCO, measures and activities conducted in or through cyberspace, outside of one’s own cyberspace, against ongoing or imminent threats to preserve freedom of action.” Source: National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-3.

<sup>354</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-3.

## Key Requirements and Activities

DCOs necessitate key requirements to be planned and conducted efficiently. The first requirement is having comprehensive and real-time cyber intelligence about adversaries' intent, capacity, and capabilities.<sup>355</sup> DCOs inherently rely on Intelligence, Surveillance and Reconnaissance (ISR) across the entire operational environment.<sup>356</sup> The second requirement is an accurate identification of friendly cyber key terrain. The third requirement is the identification and prioritization of friendly vulnerabilities.<sup>357</sup> These requirements will allow the commander to focus defensive efforts on cyber key terrain, vulnerabilities, adversaries' intents, and TTP.<sup>358</sup> Once these requirements are accomplished, the commandant plans and executes specific DCO activities to achieve its mission. The *JDN 2017-02 Cyber Operations* lists the following activities: “perimeter defence, defence-in-depth, moving-target defence (MTD), and deceptive defence.”<sup>359</sup> Additional activities not listed in the current doctrine can be done under DCO, such as threat hunting, penetration testing (red teaming), and incident response. Table 4.1 resumes the main activities that can be done during the planning and the conduct of DCO; however, this list is not comprehensive since it is mostly based on the current doctrine.

**Table 4.1 – Defensive Cyber Operations' Activities for Planning and Execution of DCO**

Planning DCO	Execution of DCO
Cyber Threat Intelligence	Perimeter Defence
Cyber Key Terrain Identification	Defence-in-Depth
Prioritization of Friendly Vulnerabilities	Moving-Target-Defence (MTD)
	Deceptive Defence
	Threat Hunting
	Penetration Testing / Red Teaming
	Incident Response

Source: National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-7 - 4-9.

Understanding the definition, key requirements, and activities is necessary to determine how AI can enhance the planning and execution of DCOs. The next section will discuss key considerations of AI for DCOs.

<sup>355</sup> *Ibid.*, 4-7.

<sup>356</sup> *Ibid.*, 4-8.

<sup>357</sup> *Ibid.*, 4-7.

<sup>358</sup> *Ibid.*

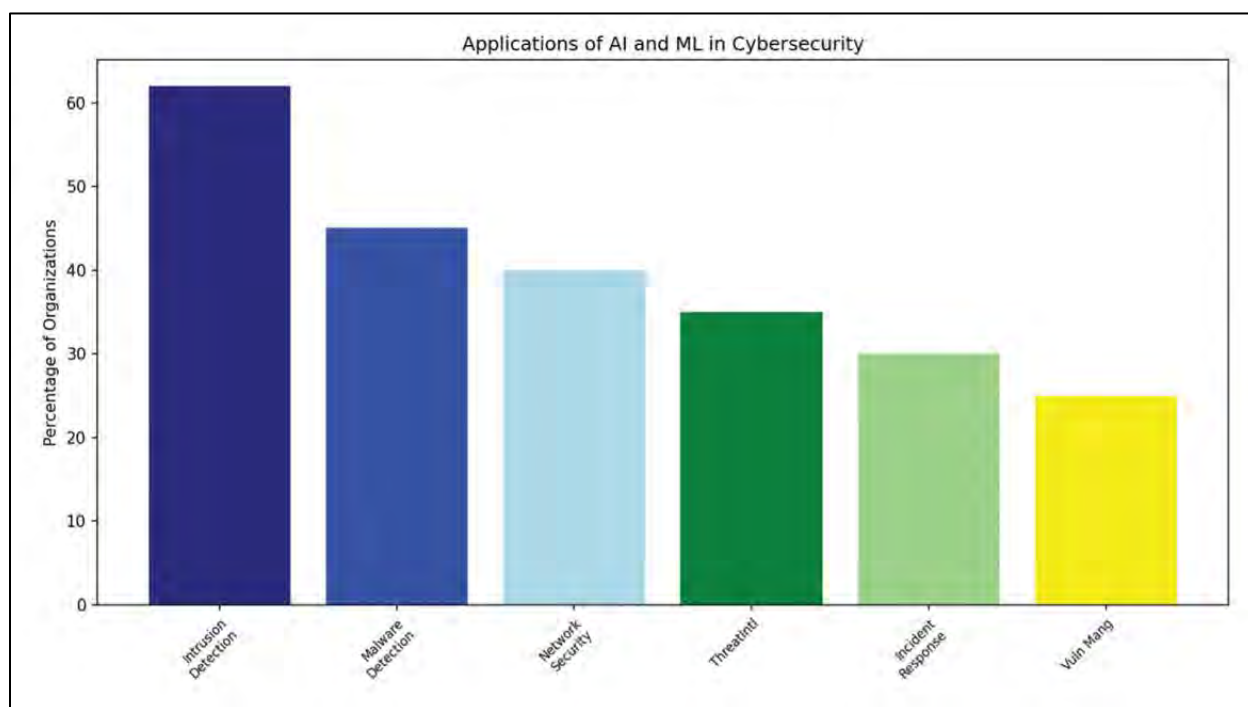
<sup>359</sup> *Ibid.*, 4-9.

## Section 4.2 – Key Considerations When Leveraging AI for DCO

It will be advisable to highlight key considerations about AI and DCOs before discussing solutions. Some factors must be considered before building an in-house AI model or using COTS solutions, such as knowing the common trends and common challenges of AI in cybersecurity.

### Current Trends of AI in Cybersecurity

The current trends in cybersecurity are leveraging AI to improve intrusion detection systems, malware detection, and network security.<sup>360</sup> Another growing and promising trend of AI is in security automation, such as vulnerability assessment, patching, and incident response.<sup>361</sup> Figure 4.1 shows the trends of AI in cybersecurity, where intrusion detection is the most popular application.



**Figure 4.1 – Application of AI in Cybersecurity**

Source: Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 24.

Currently, discriminative AI is often used in cybersecurity, especially to detect malicious traffic or malware. Expert systems are also commonly deployed as AI methods for cybersecurity where

<sup>360</sup> Nachaat Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," *Cogent Engineering* 10, no. 2 (October 25, 2023), 2. doi:10.1080/23311916.2023.2272358. <https://doi.org/10.1080/23311916.2023.2272358>.

<sup>361</sup> *Ibid.*, 9.

decision-making is needed.<sup>362</sup> These trends in cyber security can be equally applied to DCO. However, some challenges still forbid the full leverage of AI in these cybersecurity fields.<sup>363</sup> The most common challenges to employing AI in cyber security are data and the explainability of AI models.<sup>364</sup>

## Challenges

As mentioned in Chapter 2, data are at the heart of all AI models. The forthcoming lines will explain three common challenges related to data. First, leveraging AI for DCO will require significant data to create efficient models. Second, labelling data is a difficult task that requires many resources. Third, imbalanced datasets are standard in cybersecurity. Then, the challenge concerning the explainability of AI models will be discussed.

### *DCO Requires Large Datasets*

The efficiency of an AI model is primarily attributable to the quality and quantity of data it uses during its training.<sup>365</sup> The issue is that datasets used to build AI models could be very large, as discussed in Chapter 2 and demonstrated by Figure 3.2 Example of Dataset's Sizes. For instance, the Google Translate dataset contains trillions of examples.<sup>366</sup> Therefore, reviewing all examples and assessing their quality is very difficult.<sup>367</sup> Moreover, DCOs must manage large amounts of complex and mixed systems that generate large amounts of heterogeneous raw data, such as logs, reports, or alerts. Therefore, Step 2C—Preprocessing from the 10-step framework will require significant time to complete. Another challenge with data coming from different sources is that the definition of “high-risk event” or “malicious” could vary.<sup>368</sup> This could cause unexpected outcomes from the model.<sup>369</sup> Nevertheless, generating in-house data can provide opportunities to develop AI models.<sup>370</sup>

### *Labelling Data*

Another challenge related to DCO is labelling data. Leveraging AI for threat detection will require a large amount of labelled data for supervised ML algorithms.<sup>371</sup> Labelling these amounts of data manually is costly in terms of human resources, time, and finances. Semi-supervised learning techniques could be a solution where not all data must be labelled.<sup>372</sup> A good

---

<sup>362</sup> Das, "Artificial Intelligence in Cyber Security," 6.

<sup>363</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 9.

<sup>364</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 9.

<sup>365</sup> Mehta, "Explainable AI - Foundations, Methodologies and Applications," 28.

<sup>366</sup> "Data Prep - Construct Your Dataset."

<sup>367</sup> Mehta, "Explainable AI - Foundations, Methodologies and Applications," 28.

<sup>368</sup> Zhang, "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities," 1042.

<sup>369</sup> *Ibid.*

<sup>370</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 13.

<sup>371</sup> *Ibid.*

<sup>372</sup> *Ibid.*, 14.

example is an AI model trained to detect botnets that obtained an F1-score <sup>373</sup>of 0.83 with 2,400 examples labelled, whereas a similar model obtained an F1-score of 0.95 with millions of examples labelled.<sup>374</sup> The number of examples labelled between these two models is significant, with the F1-score difference being only 0.12. Despite having an F1-score lower of 0.12, the first model required only 2,400 examples labelled. This could be applied to active learning and lifelong learning principles where the model can be trained once deployed.<sup>375</sup> For instance, some deployed models improved their performance from 57% to 70% after being retrained on 700 samples labelled by a human.<sup>376</sup> Ultimately, leveraging AI for DCO will require important labelled data. However, using semi-supervised techniques and retraining the model with a small number of labelled data can reduce the downsides of using only labelled data while maintaining a high efficiency from the model.

### *Imbalanced Datasets*

DCOs require detecting sophisticated attacks from state-sponsored, inside threats, hackers, hacktivists, terrorists, and criminal organizations in real-time.<sup>377</sup> The issue is that there are very few attack examples in datasets. Since there are very few attacks compared to normal examples in datasets, it will create imbalanced data; in other words, the classification of the data set will contain uneven class proportions.<sup>378</sup> Non-malicious events will form the majority class, while malicious events will form the minority class. To illustrate this common issue, Figure 4.2 below shows the imbalanced data between attacks and normal behaviour in the CIC-IDS2017 dataset created by the Canadian Institute of Cybersecurity and the University of New Brunswick.<sup>379</sup>

---

<sup>373</sup> F1-Score is a validation metric.

<sup>374</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 14.

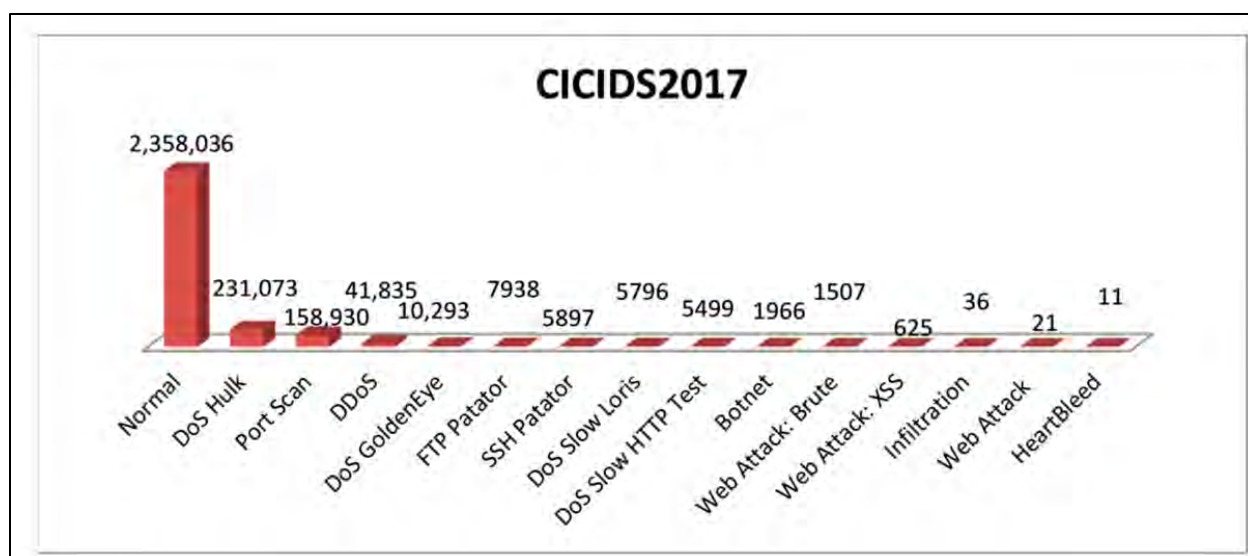
<sup>375</sup> *Ibid.*

<sup>376</sup> *Ibid.*

<sup>377</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 3-10 - 3-11.

<sup>378</sup> "Data Prep - Construct Your Dataset."

<sup>379</sup> "Datasets," last modified n.d., accessed March 6, 2024, <https://www.unb.ca/cic/datasets/index.html>.



**Figure 4.2 – Statistics of Attacks and Normal Behaviour in the CIC-IDS2017 Dataset**

Source: Ahmed Alshaibi et al., "The Comparison of Cybersecurity Datasets," *Data* 7, no. 2 (January 29, 2022), 10. doi:10.3390/data7020022. <https://doi.org/10.3390/data7020022>.

Trained on imbalanced data, the model will be biased towards the majority class and have difficulties detecting the minority class. The degree of imbalance could vary. Figure 4.3 shows the difference between mild, moderate, and extreme imbalance based on the proportion of minority class.

Degree of imbalance	Proportion of Minority Class
Mild	20-40% of the data set
Moderate	1-20% of the data set
Extreme	<1% of the data set

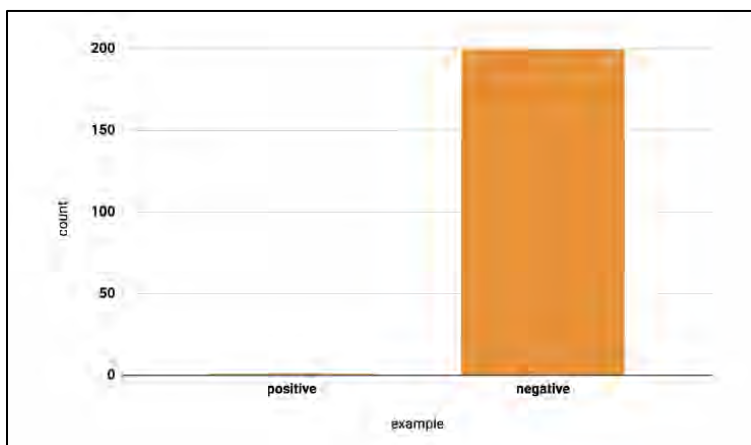
**Figure 4.3 – Degree of Imbalance based on the Proportion of Minority Class**

Source: "Data Prep - Imbalanced Data," last modified June 9, accessed April 13, 2024, <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>.

For instance, if the dataset contains less than one percent of malicious examples, it could be considered an extreme imbalance.<sup>380</sup> The problem with extreme imbalance is that the model will spend most of its training on non-malicious examples (negative) and will not learn enough

<sup>380</sup> "Data Prep - Imbalanced Data."

about malicious examples (positive).<sup>381</sup> Figure 4.4 shows the extreme imbalance between the positive and negative instances.



**Figure 4.4 – Example of Extreme Imbalanced with a Distribution of 200 Negative Instances and 1 Positive Instance**

Source: "Data Prep - Imbalanced Data."

There is a small amount of data about malicious events in datasets. Therefore, it is difficult for the model to recognize an attack or malicious behaviour on real-time data when trained on very few examples. The more examples a model sees, the better it will be capable of recognizing them.

To reduce the imbalance in the datasets, it is recommended that the majority class be downsampled and upweighed as per Figure 4.5. The first step is downsampling, which consists of “training on a disproportionately low subset of the majority examples.”<sup>382</sup> In the example above, with 1 positive to 200 negative instances (0,5%), downsampling by a factor of 10 will improve the imbalance of 1 to 20 negatives (5%).<sup>383</sup> The second step, upweighting, means “adding an example weight to the downsampled class equal the factor by which you downsampled.”<sup>384</sup> In this example, weights will be added to the downsampled instances.<sup>385</sup>

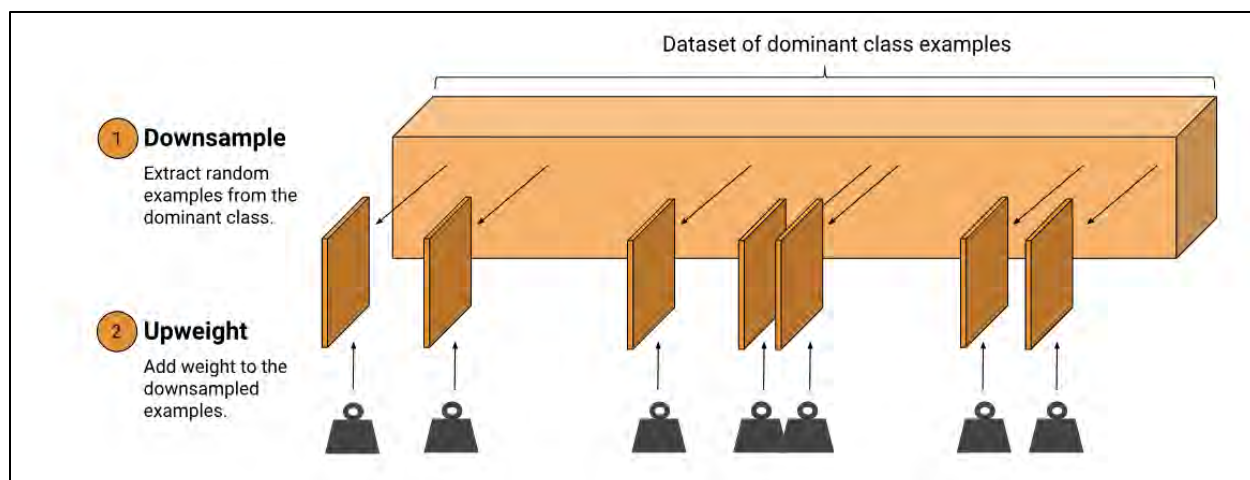
<sup>381</sup> *Ibid.*

<sup>382</sup> *Ibid.*

<sup>383</sup> *Ibid.*

<sup>384</sup> *Ibid.*

<sup>385</sup> *Ibid.*



**Figure 4.5 – Steps of Downsampling and Upweighting**

Source: "Data Prep - Imbalanced Data."

Upweighting the majority class seems counterintuitive since the goal is to improve the model on the minority class; however, this method will allow faster convergence, improve disk space, and ensure the model is still calibrated.<sup>386</sup> Another challenge to bear in mind is that the low number of malicious example instances available in the datasets will need to be split between training, validation, and testing datasets.

#### *AI Models Interpretability and Explainability*

Another common challenge for DCO is AI models' lack of interpretability and explainability.<sup>387</sup> DCO may require transparent and interpretable outputs. Threat modelling could be a good example that would demand explanations of the model output. As explained in Chapter 3, not all AI models are interpretable or explainable. AI model outputs cannot be interpretable or explainable when unsupervised learning algorithms and DL models are used; however, the output for shallow ML with supervised algorithms is explainable. Therefore, if it is critical for the organization to explain output, it would be recommended to use supervised algorithms. Nevertheless, this would limit the organization since DL and unsupervised algorithms can produce efficient AI models.

### **Section 4.3 – AI for Planning DCO**

This section will propose solutions that leverage AI to improve the CAF Cyber Force's DCO planning. As mentioned before, DCOs are threat—and mission-oriented. Therefore, DCO planning relies heavily on intelligence to identify potential threats and prioritize key terrains.<sup>388</sup>

<sup>386</sup> *Ibid.*

<sup>387</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 5.

<sup>388</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-6 - 4-7.

The section will propose solutions for enhancing intelligence and facilitating cyber key terrain identification to improve DCO planning efforts.

### Cyber Threat Intelligence

Cyber threat intelligence is done through the collection and analysis of data to predict future cyber-attacks.<sup>389</sup> DCOs cannot exist without accurate and reliable intelligence and threat analysis. ISR enables DCOs to sense, detect, and manoeuvre against adversaries. The cyberspace landscape has evolved significantly and is becoming increasingly complex. Moreover, the exponential growth of data has made manual and reactive approaches outdated and insufficient for the current threat landscape.<sup>390</sup> Therefore, human involvement in analyzing all data and producing intelligence reports about adversaries' cyberspace activities is no longer viable.<sup>391</sup>

Leveraging AI to manage threat intelligence could significantly improve the efficiency and speed of DCOs.<sup>392</sup> First, AI could automate threat intelligence processes such as data collection and analysis, freeing analysts to complete more complex tasks.<sup>393</sup> Threat intelligence will become more efficient since AI models can be improved and trained constantly.<sup>394</sup> Second, a proactive approach, also called threat prediction, could be used where AI augments threat intelligence and then provides models to forecast potential threats and vulnerabilities of the organization.<sup>395</sup> Improving threat intelligence with AI will enhance the overall cybersecurity and the conduct of DCOs.<sup>396</sup> Finally, AI models could improve threat intelligence sharing and facilitate its dissemination with allies and partners.<sup>397</sup> These solutions would improve the overall access to threat intelligence and, therefore, improve the defence posture against potential threats.<sup>398</sup>

### AI Solutions

To ameliorate data analysis and threat intelligence for DCOs, the CAF Cyber Force could leverage Swarm intelligence, which belongs to the Bio-inspired Computation AI subfield described in Chapter 2. Swarm intelligence algorithms mimic the behaviour of swarm insects, fishes, and birds' behaviour.<sup>399</sup> These swarms are inspiring since, individually, they cannot solve

---

<sup>389</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 14.

<sup>390</sup> Siva Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence" Taylors University Subang Jaya Malaysia, 2023), 3.

<sup>391</sup> *Ibid.*

<sup>392</sup> *Ibid.*, 4.

<sup>393</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 9.

<sup>394</sup> Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," 3.

<sup>395</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 10.

<sup>396</sup> Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," 4.

<sup>397</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 10.

<sup>398</sup> *Ibid.*

<sup>399</sup> Pradeep Kumar and B. R. Prasad Babu, "Investigating Open Issues in Swarm Intelligence for Mitigating Security Threats in MANET," *International Journal of Electrical & Computer Engineering* 5, no. 5 (October, 2015): 1197. <http://iaesjournal.com/online/index.php/IJECE/issue/archive>.

complex problems; however, by interacting with each other, they will accomplish difficult tasks such as finding the shortest route for food, building nests or travelling as a single and coherent entity without the involvement of a centralized authority.<sup>400</sup> Swarm intelligence algorithms are often employed for threat detection and response.<sup>401</sup> They can facilitate sharing threat intelligence, analyzing data, and making real-time decisions.<sup>402</sup>

### Cyber Key Terrain

Defensive land operations require the commander to identify vital ground and key terrain.<sup>403</sup> Key terrain is defined as “any locality, or area, the seizure or retention of which affords a marked advantage to either combatant.”<sup>404</sup> A vital ground is “a ground of such importance that it must be retained or controlled for the success of the mission.”<sup>405</sup> In defensive operations, if the vital ground is lost in the hands of the adversary, the defence is untenable.<sup>406</sup> Once the commander has identified its vital ground, avenues of approach toward it will also be identified.<sup>407</sup> For each avenue of approach, commanders will establish grounds that offer a marked advantage, which should be held to establish its defensive position.<sup>408</sup> These grounds are the key terrain, and usually, troops are tasked to hold these terrains.

These concepts from land operations also apply to cyber operations. Since DCOs are mission-oriented, it is necessary to identify cyber key terrain (CKT), which is an essential planning component.<sup>409</sup> The *JDN 2017-02 Cyber Operations* does not formally define CKT since it is a new concept.<sup>410</sup> CKT defines a connection or a node in the network as valuable and critical for both the friendly and adversary forces.<sup>411</sup> Identifying CKTs for DCO can be very difficult since the cyber domain is very complex. CKT can be in all cyber domain layers, such as the physical network layer, logical network layer, and/or cyber persona layer; see Figure 4.2 for more details.<sup>412</sup>

---

<sup>400</sup> *Ibid.*

<sup>401</sup> Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," 29.

<sup>402</sup> *Ibid.*

<sup>403</sup> National Defence, *B-GL-300-001-FP-001 — Land Operations* (Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2008), 7-48.

<sup>404</sup> *Ibid.*

<sup>405</sup> *Ibid.*

<sup>406</sup> *Ibid.*

<sup>407</sup> *Ibid.*

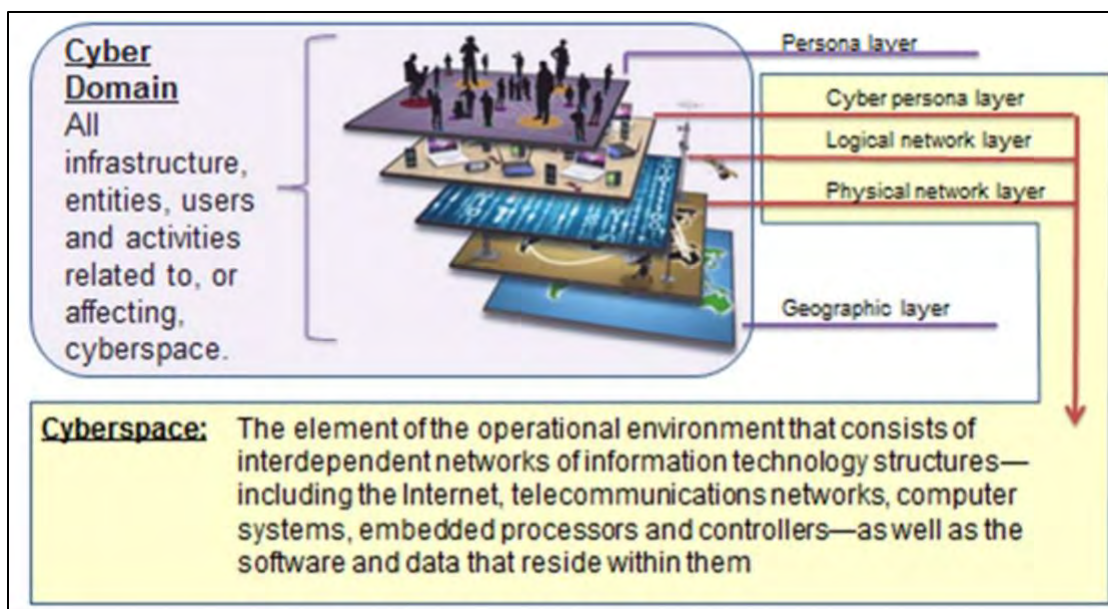
<sup>408</sup> *Ibid.*

<sup>409</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 5-11.

<sup>410</sup> *Ibid.*, 4-7.

<sup>411</sup> Longhui Liu et al., "Improved Technique for Order of Preference by Similarity to Ideal Solution Method for Identifying Key Terrain in Cyberspace Asset Layer," *Plos One* 18, no. 7 (July 13, 2023): 2. doi:10.1371/journal.pone.0288293. 288293. <https://doi.org/10.1371/journal.pone.0288293>

<sup>412</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 5-11.



**Figure 4.6 – The Cyber Domain Layers**

Source: National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 2-2.

### *AI Solutions*

After reviewing the literature, nothing was found about AI solutions to facilitate the identification of CKT. Two articles were reviewed but did not mention AI. The first article published in 2023, *Improved technique for order of preference by similarity to ideal solution method for identifying key terrain in cyberspace asset layer*, proposes techniques to improve the precision of CKT identification and to understand the relationship between CKT and cyber attack mission.<sup>413</sup> The second article, also published in 2023, *Cyberspace Knowledge Base for Key Terrain Identification*, proposes the model Cyberspace Knowledge Base to improve the identification of key terrain.<sup>414</sup> Due to the lack of publication about AI to improve the identification and management of CKT, no recommendations will be made for this activity in DCO planning.

### **Section 4.4 – AI for Executing DCO**

The DCO definition provided above indicates crucial tasks that need to be accomplished: “detect, defeat and/or mitigate offensive and exploitive actions.”<sup>415</sup> These tasks could be divided

<sup>413</sup> Liu, "Improved Technique for Order of Preference by Similarity to Ideal Solution Method for Identifying Key Terrain in Cyberspace Asset Layer," 1.

<sup>414</sup> Kaiqi Yao et al., "Cyberspace Knowledge Base for Key Terrain Identification," *Electrical and Electronics Engineers (IEEE)* (August 21, 2023), 728. doi:10.1109/AINIT59027.2023.10212951. <https://ieeexplore-ieee-org.cfc.idm.oclc.org/document/10212951>.

<sup>415</sup> National Defence, *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations*, 4-3.

into three categories: prevention, detection, and defeat. This section will discuss how CAF Cyber Command can leverage AI to accomplish these three DCO activities.

## Prevention

### *Vulnerability Assessment*

Vulnerabilities must be identified, rated, and prioritized to defend one's networks and systems. This activity is critical since adversaries will attempt to exploit vulnerabilities.<sup>416</sup> The current vulnerability assessment (VA) methods mostly involve manual inspection of network architecture, system configurations, software codes, or physical installation.<sup>417</sup> Most of the time, vulnerability assessments are inefficient since they require significant human efforts, are time-consuming, not easily scalable, and are not comprehensive.<sup>418</sup> Some parts of VA can be done automatically but will require human assessment and actions once the assessment is completed. Furthermore, VA must be done regularly. Consequently, it is reasonable to assume that cybersecurity personnel spend a significant portion of their time conducting VA. Moreover, since they are mostly done manually, important portions of the organization's systems might not be assessed properly. Therefore, having a full portrait of the organization's vulnerabilities is almost impossible. This is a daunting task since keeping pace with the constantly involving threat landscape is almost impossible.<sup>419</sup>

### *AI Solutions*

Leveraging AI could revolutionize VA by offering automated, scalable, and proactive solutions to reduce the burden on cyber security personnel and counter the new AI-enabled intrusion attempts.<sup>420</sup> Several organizations offer solutions with AI-enabled VA.<sup>421</sup> An assessment should be done to determine if the CAF should procure COTS tools available in the industry or build an in-house AI-enabled VA model. Different types of AI can be used for vulnerability assessment and management. First, expert systems and rule-based systems are common AI techniques that automate security processes, such as vulnerability assessment and management.<sup>422</sup> Second, Gen AI with NLP models could be used to improve vulnerability assessment by analyzing security advisories.<sup>423</sup> NLP can potentially be used to identify important information from the organization's advisories, and established the severity of the

---

<sup>416</sup> Adewale Daniel Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1725.

<sup>417</sup> *Ibid.*

<sup>418</sup> *Ibid.*

<sup>419</sup> *Ibid.*

<sup>420</sup> *Ibid.*, 1726.

<sup>421</sup> *Ibid.*

<sup>422</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 8.

<sup>423</sup> *Ibid.*, 17.

vulnerability.<sup>424</sup> Third, an AI model built with reinforcement learning can be used to patch vulnerabilities when discovered by the AI model.<sup>425</sup> Finally, AI models can be leveraged to continuously analyze the organization's networks and systems to discover and patch vulnerabilities.<sup>426</sup> However, current AI-based vulnerability management solutions will not cover all vulnerabilities. It is still recommended that penetration testing and security audits be conducted.<sup>427</sup>

## Detection

### *Intrusion Detection System*

Intrusion Detection Systems (IDS) are essential for detecting abnormal activities or intrusions in the CAF cyberspace component. Traditional IDS uses two detection methods: signature-based detection and rule-based detection. Signature-based detection relies on known threats, malware, or attack techniques.<sup>428</sup> Rule-based detection relies on rules or heuristics to detect suspicious activities or behaviours.<sup>429</sup> The current problem with IDS is that it cannot effectively detect new threats or different attack patterns.<sup>430</sup> This problem could be countered by leveraging AI-based IDS.<sup>431</sup>

### *AI Solutions*

The most common and promising application of AI in cyber security or DCO is with IDS.<sup>432</sup> AI can significantly improve traditional IDS by learning from large data volumes of network traffic, log data, and security event telemetry in real-time.<sup>433</sup> Moreover, AI improves the contextualization and correlation of data and can help identify subtle indicators of compromise from a sophisticated intrusion.<sup>434</sup> Both machine learning and deep learning have been leveraged to create an AI-based solution for IDS to continuously adapt to evolving threats, identify unseen attacks and detect sophisticated attacks that can evade traditional IDS.<sup>435</sup> Table 4.2 summarizes the application of shallow ML and DL AI subfields that could be used to develop AI models. The

---

<sup>424</sup> *Ibid.*

<sup>425</sup> *Ibid.*

<sup>426</sup> *Ibid.*

<sup>427</sup> *Ibid.*, 18.

<sup>428</sup> Adewale Daniel Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1722.

<sup>429</sup> *Ibid.*

<sup>430</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 5.

<sup>431</sup> *Ibid.*

<sup>432</sup> *Ibid.*, 4.

<sup>433</sup> Adewale Daniel Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1721.

<sup>434</sup> *Ibid.*

<sup>435</sup> *Ibid.*, 1724.

recommended supervised algorithms are decision trees and support vector machines.<sup>436</sup> Unsupervised algorithms could be useful to improve the detection of unknown threats and malware.<sup>437</sup> DL algorithms can also be used to improve the detection of unknown threats and unusual behaviour.<sup>438</sup> AI-enabled IDS would mostly use discriminative AI since IDS requires classification solutions. Although generative AI can also be leveraged to augment the resilience of the model during the training and testing steps with a Generative Adversarial Networks algorithm that can generate synthetic data.<sup>439</sup>

**Table 4.2 – Possible AI Solutions to Improve Intrusion Detection System**

Objective / Application to IDS	AI Subfield	Training Algorithm and Technique	Description
Improve identification of known threats and malware.	Discriminative – Classification – Shallow Machine Learning	Supervised Learning	The model is trained on labelled data for classification, such as malicious versus benign.
Improve identification of unknown threats and unusual behaviour.		Unsupervised Learning	The model is trained to identify patterns among unlabeled data to detect anomalies.
Improve response strategies to sophisticated and evolving threats and attacks.		Reinforcement Learning	The model is trained through trial and error to optimize decisions. This method does not provide labelled data. The ML model received the current state, specific goal, and list of allowable actions and constraints for the outcome. <sup>440</sup>
Improve identification of malicious patterns in network traffic and malware binaries.	Discriminative – Classification – Deep Learning	Convolutional Neural Network (CNN)	This algorithm is usually employed for computer vision and speech recognition tasks. It is also good at analyzing structured data, such as images.

<sup>436</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 4.

<sup>437</sup> Adewale Daniel Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1724.

<sup>438</sup> *Ibid.*

<sup>439</sup> *Ibid.*

<sup>440</sup> Janiesch, "Machine Learning and Deep Learning," 687.

Improve identification of anomalies from the system behaviour.		Recurrent Neural Networks	This algorithm is used for sequential data structures, such as time-series data, event sequences or natural language.
Improve the model robustness during training and testing.	Generative AI	Generative Adversarial Networks (GAN)	GAN will generate synthetic data and adversarial examples to improve the model during the testing step.

Sources: Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1724.

Janiesch, "Machine Learning and Deep Learning," 687.

Figure 4.7 summarizes the pros and cons of using supervised versus unsupervised approaches for threat detection. Shallow ML-supervised algorithms can be used for the entire AI model but require labelling the data, which requires a lot of human, financial, and time resources.<sup>441</sup> Conversely, shallow ML unsupervised algorithms do not require labelled data but are only used as an addition to the AI model for detection.<sup>442</sup>



**Figure 4.7 – Pros and Cons of Supervised and Unsupervised ML for Cyber Threat Detection**

Source: Apruzzese, "The Role of Machine Learning in Cybersecurity," 7.

### *Malware Detection*

Another component of DCO is detecting malware, which is “malicious software that can harm computers.”<sup>443</sup> Cyber malicious actors have used malware over the past few decades, and they are becoming more sophisticated, such as polymorphic and metamorphic malware.<sup>444</sup> Attackers can bypass traditional malware detection, which relies on existing rules and signatures.

<sup>441</sup> Apruzzese, "The Role of Machine Learning in Cybersecurity," 7.

<sup>442</sup> *Ibid.*

<sup>443</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 4.

<sup>444</sup> *Ibid.*

## AI Solutions

Leveraging AI solutions to improve malware detection will help detect new threats and cyber-attack patterns.<sup>445</sup> Both shallow ML and DL are being used to improve malware detection. Common training algorithms for malware detection are “decision trees, neural networks, and support vector machines.”<sup>446</sup> Table 4.3 proposes techniques to improve malware identification and categorization.

**Table 4.3 – Possible AI Solutions to Improve Malware Detection**

Objective / Application to Threat Hunting	AI Subfield	Training Algorithm and Technique	Description
Improve identification and categorization of malware.	Discriminative – Classification – Deep Learning	Convolution Neural Networks (CNNs) Recurrent Neural Networks (RNNs)	The CNN algorithm is often used for computer vision, speech recognition, images, and videos. The RNN algorithm is often employed for sequential data such as time series or natural language.
Enhance the AI model training capabilities. <sup>447</sup>	Generative	GAN Variational Autoencoders (VAEs)	See Table 2.5

Sources: Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 6.

Al-Dosari, "Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges," 690.

For malware detection, it is recommended to use an ensemble architecture that combines the knowledge of several models and significantly improves the performance and reliability of the malware detection model.<sup>448</sup>

<sup>445</sup> *Ibid.*

<sup>446</sup> *Ibid.*

<sup>447</sup> *Ibid.*, 6.

<sup>448</sup> *Ibid.*

## *Threat Hunting*

The saying “Looking for a Needle in a Haystack” helps to imagine what threat hunting is. Cyber operators must find anomalies, malware, or indicators of compromise in a huge quantity of data. Traditional malware and IDS cannot detect all sophisticated attacks, especially those from APT. Threat hunting is “any manual or machine-assisted process for finding security incidents that your automated detection systems missed.”<sup>449</sup> Another definition of threat hunting is “the procedure of searching across networks and endpoints for threats that have eluded security measures and are about to launch an attack or accomplish their objectives.”<sup>450</sup> Threat hunting requires human involvement, which could vary based on the organization's maturity.<sup>451</sup>

## *AI Solutions*

Leveraging AI can certainly improve threat hunting by automating and augmenting the speed of analyzing large amounts of data, identifying patterns, and detecting threats or intrusions. The most common AI techniques used in threat hunting are NLP for text analysis of a large amount of data, DL to recognize patterns in data, and graph analysis to identify relationships between data and networks.<sup>452</sup>

## *Defeat*

## *Incident Response*

Identification and triage, containment, remediation and recovery, post-incident review and lessons learned are common activities in traditional incident response.<sup>453</sup> Nowadays, these common steps are part of several frameworks and are used by many organizations to respond to cyber incidents or cyber-attacks. However, these methods have several downsides. They are intensively time-consuming, require human resources, are limited in scalability, and are prone to human error in manual processes.<sup>454</sup>

AI could provide solutions to counter most incident response's downsides by automating most if not all, processes, such as triaging security alerts and reducing the time it takes to detect

---

<sup>449</sup> splunk, *The PEAK Threat Hunting Framework - Modernized Hunting for the Evolving Threat Landscape* (United States: Splunk Inc., 2023), 4.

<sup>450</sup> Ali Aghamohammadpour, Ebrahim Mahdipour and Iman Attarzadeh, "Architecting Threat Hunting System Based on the DODAF Framework," *The Journal of Supercomputing* 79, no. 4 (September 23, 2022), 4216. doi:10.1007/s11227-022-04808-6. <https://doi.org/10.1007/s11227-022-04808-6>.

<sup>451</sup> splunk, *The PEAK Threat Hunting Framework - Modernized Hunting for the Evolving Threat Landscape*, 4.

<sup>452</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 12.

<sup>453</sup> Sontan, "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities," 1727.

<sup>454</sup> *Ibid.*

an intrusion.<sup>455</sup> AI solutions in incident response are also scalable. They could also improve the organization's situational awareness by analyzing historical incident data and threat patterns.<sup>456</sup>

### *AI Solutions*

For automated tasks, expert systems and rule-based systems can replicate human decision-making by automating tasks and decisions from predefined rules.<sup>457</sup> Currently, rule-based systems are a common AI technique used to automate cybersecurity processes and tasks, such as incident response.<sup>458</sup> This type of AI model can also be implemented with IDS and security information and event management (SIEM) systems to improve the organization's incident response process.<sup>459</sup> AI incident response models also use the decision tree supervised algorithm to help automate decision-making and classify security as benign or malicious.<sup>460</sup> Table 4.4 summarizes possible methods to leverage AI to improve incident response. AI models in incident response still need improvement to be more interpretable and explainable and require reducing the number of false positives.<sup>461</sup>

**Table 4.4 – Possible AI Solutions to Improve Incident Response**

Objective / Application to Threat Hunting	AI Subfield	Training Algorithm and Technique	Description
Minimize response time by automated actions on compromised systems, such as containment <sup>462</sup>	N/A	Rule-based systems	A technique that uses predefined rules and algorithms to deal with events. <sup>463</sup>
		Expert Systems	A technique that combines rule-based and knowledge-based systems to make decisions automatically on complex security tasks.

<sup>455</sup> *Ibid.*, 1728.

<sup>456</sup> *Ibid.*

<sup>457</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 10.

<sup>458</sup> *Ibid.*, 8.

<sup>459</sup> *Ibid.*, 9.

<sup>460</sup> *Ibid.*

<sup>461</sup> *Ibid.*

<sup>462</sup> Raja Sindiramutty, "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence," 2.

<sup>463</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 8.

Automate the decision-making process for incident response by classifying security events as benign or malicious. <sup>464</sup>	Discriminative – Classification – Shallow ML	Supervised algorithm: Decision trees	This algorithm can be applied to regression and classification problems. It classifies the data into small groups until these groups cannot be divided further. The algorithm uses the nodes and leaves of the tree. The nodes are decisions about data, and the leaves are the data categorized. <sup>465</sup>
--	--	--	--

Source: Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 9-10.

## Summary

This chapter intends to provide AI solutions and recommendations to improve the CAF Cyber Command's planning and conduct of DCOs. To accomplish this objective, this chapter was divided into four sections. The first section defined DCO and identified its key requirements and activities. The second section presented the current trends of AI in cybersecurity and the challenges when leveraging AI, which are related to data and explainability of AI models. The third section proposed solutions to improve the planning of DCO by integrating AI in cyber threat intelligence and identification of cyber key terrain. However, this study could not propose AI solutions for the identification of cyber key terrain. This is an area that might be developed in the future. The last section proposed solutions to improve the conduct of DCO by preventing, detecting, and defeating cyber malicious actors. This study did not cover all possible applications in cybersecurity that could leverage AI.

<sup>464</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 9.

<sup>465</sup> "Generative Vs. Discriminative Machine Learning Models."

## CHAPTER 5 - CONCLUSION AND RECOMMENDATIONS

This study aimed to answer the question: How can the CAF Cyber Command leverage AI to conduct Defensive Cyber Operations (DCOs) and become AI-enabled by 2030? The proposed thesis was the following: The CAF Cyber Command must leverage AI to improve the efficiency of key planning and conducting DCO activities by following a framework to create AI models while complying with the laws, regulations, and policies of the Government of Canada and DND/CAF and aligning with the vision of *The DND and CAF AI Strategy*. These recommendations will support the CAF Cyber Command's objective of becoming an AI-enabled organization by 2030, which supports the ambitious commitment set by *The DND and CAF AI Strategy*'s vision.

This study was divided into five chapters. The first chapter discusses the importance of becoming AI-enabled by 2030 and how AI can improve DCOs. The second chapter explained technical AI concepts and model layers and components. The third chapter proposed a framework for building AI models and concerns the CAF Cyber Command leadership's responsibilities when leveraging AI. It explained the risks associated with AI, described the regulations and compliance that must be followed when leveraging AI, and presented an emergent field in AI that will help reduce some risks. The fourth chapter explained how AI can be leveraged for the planning and conducting of DCOs. Finally, this chapter presents recommendations and conclusions about this study.

### Section 5.1 – 8 Recommendations

Eight recommendations can be made from the content presented in the previous chapters. Applying these recommendations will contribute to the CAF Cyber Command becoming an AI-enabled organization by 2030 and answer how AI can be leveraged for DCO.

#### Recommendation 1 – Must become AI-enabled Organization Soon

The first recommendation is that the CAF Cyber Command acknowledge the urgency of becoming an AI-enabled organization. Commitment and effort should be invested in developing AI-enabled systems to improve cyber operations. Nevertheless, AI is not the solution to all problems. As mentioned in the *PFEC*, leveraging AI requires a deliberate and thoughtful approach to comprehensively frame the problem.<sup>466</sup> In other words, the CAF Cyber Command must become AI-enabled to counter its adversaries in cyberspace; however, a comprehensive assessment must be done before initiating any AI project and it must be done as soon as possible.

---

<sup>466</sup> National Defence, *Pan-Domain Force Employment Concept - Prevailing in a Dangerous World*, 26.

## Recommendation 2 – A 10-Step Framework

If the CAF Cyber Command leadership decides to create in-house AI models, it is recommended that they follow a framework. Before using this framework, it is recommended to assess what is currently done by the industry. To recap, Chapter 3 proposed a 10-step framework for creating, deploying, and defending AI models. This framework is a simple and general guideline and does not include all technical details. The 10-step framework helps frame the problem properly, identify and prepare datasets, and select the appropriate architecture model and learning algorithms. It also includes requirements such as infrastructure selection, validating, testing the model, deploying, monitoring, and defending the model. The 10-step framework should be adapted to the situation and needs of the organization. Additionally, it is important to document all steps when creating an AI model. Table 5.1 provides a summary of the 10-step framework to build AI models, which was presented in Chapter 3.

**Table 5.1 – 10 Step Framework Summary**

Steps	Description
<b>Step 1 – Framing the Problem</b>	Answer the following questions: <ul style="list-style-type: none"> <li>• What is the problem that needs to be solved?</li> <li>• What is the context?</li> <li>• What are the tasks that need to be accomplished?</li> <li>• What are the objectives to achieve?</li> <li>• What kind of output is needed to solve the problem?</li> <li>• Is there a requirement to understand the internal workings of the model for transparency, accountability, and trust?</li> <li>• Does the organization have already an AI model that achieve this objective that could be used?</li> <li>• Is AI the only way to solve the problem?</li> </ul>
<b>Step 2 – Datasets</b>	This step selects and prepares datasets to train, validate and test the model.
<b>Step 2A – Raw Data Collection</b>	<ul style="list-style-type: none"> <li>• Collecting raw data to build an in-house dataset.</li> <li>• Identifying public datasets that can be used.</li> </ul>
<b>Step 2B – Identify Features and Labels</b>	<ul style="list-style-type: none"> <li>• Identifying features.</li> <li>• Labelled data (if applicable).</li> </ul>
<b>Step 2C - Preprocessing</b>	<ul style="list-style-type: none"> <li>• Cleaning the data.</li> <li>• Feature engineering.</li> </ul>
<b>Step 2D – Sampling and Split the Datasets</b>	Sampling and splitting the datasets into the training, validating, and testing sets.
<b>Step 3 – Selecting the Model</b>	This step includes three sub-steps.

<b>Step 3A – Selecting the AI Subfield</b>	Selecting the most appropriate AI subfield for the problem between expert systems, shallow ML, DL, or BIC.
<b>Step 3B – Selecting the AI Model</b>	Selecting the most appropriate model architecture for the problem between discriminative and Gen AI.
<b>Step 3C – Selecting the Training Algorithms</b>	Selecting the most appropriate training algorithms between supervised, unsupervised, reinforcement, and DL algorithms.
<b>Step 4 – Infrastructure Requirement</b>	Identifying the infrastructure requirements for the AI model.
<b>Step 5 – Training the Model</b>	Train the model with the training set.
<b>Step 6 – Validating the Model</b>	<ul style="list-style-type: none"> <li>• Validate the model with the validation set.</li> <li>• For supervised algorithms, check for overfitting and underfitting.</li> <li>• Adjust parameters and hyperparameters.</li> <li>• Complete the Algorithmic Impact Assessment</li> <li>• Complete the GBA+</li> </ul>
<b>Step 7 – Testing the Model</b>	Test the model with the testing set.
<b>Step 8 – Deploying the Model</b>	<ul style="list-style-type: none"> <li>• Deployment of the model in the production environment.</li> <li>• Publish policies related to the model.</li> </ul>
<b>Step 9 – Monitoring and Maintaining the Model</b>	Monitor, retrain, and update the model regularly.
<b>Step 10 – Defending the Model</b>	Defend the model against adversarial machine learning.

### Recommendation 3 – Reduce Output Risks

It is acknowledged that AI models, especially those using Gen AI, could make hallucinations or present biased outputs. Chapter 3 provides several solutions to reduce this risk, such as writing proper AI usage policies, selecting only the data that is relevant to the task when building datasets, reviewing outputs, using security-oriented tools and platforms, and removing any PII and classified data. Section 3.2 in Chapter 3 provides more information about mitigation measures to reduce the output risks.

### Recommendation 4 – Defending AI Models

It is strongly recommended that the defence of all DND and CAF deployed AI models and AI-enabled systems be planned deliberately. AI can be leveraged to improve DCO, and DCO can be conducted to defend DND and CAF AI-enabled systems. A few elements that should be

remembered about the defence of AI models are the following. First, there is a gap and differences between traditional cyber defence for traditional IT systems and cyber defence for AI models. Second, threat modelling on AI models should be done with the MITRE ATLAS<sup>TM</sup>.<sup>467</sup> Third, red teaming against AI models should be conducted to assess the AML vulnerabilities. Especially with Gen AI, the red team should also assess the model output risks by trying to get the model to generate harmful content.<sup>468</sup> These red teaming assessments should be done before the deployment in production and conducted regularly once the AI model is deployed. It is critical to ensure the integrity of datasets, which are critical targets for malicious cyber actors. Moreover, CAF Cyber Command leadership should assess the advantages and disadvantages of having an in-house red team or hiring external firms. Developing this expertise in-house could be valuable and aligned with LoE 3: Ethics, Safety and Trust and LoE 4: Talent and Training from *The DND and CAF AI Strategy*.<sup>469</sup> Fourth, the entire cycle (design, development, deployment, and maintenance) in the AI Model Lifecycle should be done with security in mind. Finally, it is recommended to coordinate with the Digital Transformation Officer (DTO) and, if it is created with the DCAIC, before initiating any AI-enabled project.

#### Recommendation 5 – Be Aware of Regulations and Compliance

Before initiating any AI project, the CAF Cyber Command leadership and staff should be familiar with all regulations and compliance regarding AI. As mentioned in the previous recommendation, it is recommended that the leadership and staff contact the DTO and DCAIC before beginning the project. The leadership and staff should also be familiar with *The DND and CAF AI Strategy*. Table 5.2 provides a summary of all regulations and compliance, which was presented in Chapter 3, that must be followed when creating AI models.

**Table 5.2 – Laws, Policies, Guidelines, Guidance, and Commitment for the Development and Implementation of AI-enabled Technologies**

Regulation Type	Name
Laws	Canadian Law
	International Law
Regulations and Government of Canada Policies	Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems
	Directive on Automated Decision-Making
Guidelines	Guide on the use of Generative AI
Guidance from External Review Bodies	National Security and Intelligence Committee of Parliamentarians (NSICOP)

<sup>467</sup> "Mitre Atlas."

<sup>468</sup> "How to Red Team a Gen AI Model."

<sup>469</sup> National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 18-20.

	National Security and Intelligence Review Agency (NSIRA)
	Office of the Privacy Commissioner (OPC)
Commitment	Gender-Based Analysis (GBA) Plus

Source: National Defence, *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*, 19.

#### Recommendation 6 – Explore XAI Development

If understanding the internal workings, transparency, and interpretability of the AI model is required, shallow ML and supervised training algorithms are recommended. However, these models could be less performant than more complex models, such as DL. Other solutions to obtain better transparency and interpretability of DL models could be possible with XAI. Although, this is still an emerging field, and it is not fully developed. In the coming years, XAI might propose viable solutions to improve transparency and interpretability in AI.

#### Recommendation 7 – Be Cognizant of Current Trends and Challenges of AI in DCOs

The CAF Cyber Command leadership and staff should be cognizant of current trends and challenges related to leveraging AI in cybersecurity. At the time of writing, AI is mostly used for IDS, malware detection, network security, cyber threat intelligence, incident response, and vulnerability management.<sup>470</sup> These are the common cybersecurity applications where AI is leveraged. Therefore, the CAF Cyber Command should begin leveraging AI for DCO in these applications.

The common challenges related to using AI in cybersecurity involve data and the explainability of AI models. The CAF Cyber Command leadership and staff should be cognizant that AI for DCO purposes necessitates a significant amount of data to train models efficiently. The other challenge with data is the high cost of labelling data in terms of human resources, time, and finances. Finally, when selecting public datasets or building in-house datasets, it is important to be cognizant of the risk of having imbalanced datasets, which could reduce the efficiency of the AI model in recognizing rare occurrences. Downsampling and upweighting are techniques to reduce the degree of imbalance in datasets.

#### Recommendation 8 – Leveraging AI for DCO's Activities and Managing Models

As discussed in Section 4.3 – AI for Planning DCO and Section 4.4 – AI for conducting DCO, AI can be leveraged to improve the planning and execution of DCOs. In the planning of DCO, there are opportunities to leverage AI to improve cyber threat intelligence. Conversely, the literature did not provide many AI solutions to identify cyber key terrain and vital ground.

<sup>470</sup> Mohamed, "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey," 24.

Nevertheless, with the rapid evolution of AI, solutions might be, if not already, possible soon. Regarding the prevention, detection, and defeat of cyber adversaries, several solutions for AI already exist, especially for IDS, vulnerability management, malware detection, threat hunting, and incident response. It is recommended that the CAF Cyber Command leadership starts leveraging AI in these applications.

As mentioned in the first recommendation, leveraging AI will require a lot of effort and resources. First, building AI models will be challenging since it is a new practice. Second, AI models will need to be consistently monitored, updated, and defended. As mentioned in Chapter 3, each AI-enabled system increases the attack surface. AI will improve DCO; however, it will bring additional responsibilities to the CAF Cyber Command.

## **Section 5.2 – Conclusion**

To conclude, the application of the eight recommendations outlined above provides a comprehensive roadmap for the CAF Cyber Command to leverage AI to conduct Defensive Cyber Operations (DCOs) and become AI-enabled by 2030. However, it is essential to mention that not all AI concepts, risks, challenges, and solutions could be fully addressed in this study. While this study focused on recommendations for DCO, an assessment should be done about AI to cover the full spectrum of cyber operations, including offensive cyber operations. Despite the inherent challenges and obstacles of AI, these recommendations should initiate the discussion among the CAF Cyber Command leadership and staff insofar as the discussion begins now.

## BIBLIOGRAPHY

- Adewale Daniel Sontan and Segun Victor Samuel. "The Intersection of Artificial Intelligence and Cybersecurity: Challenges and Opportunities." *World Journal of Advanced Research and Reviews* 21, no. 2 (February 28, 2024): 1720-1736. doi:10.30574/wjarr.2024.21.2.0607. <https://doi.org/10.30574/wjarr.2024.21.2.0607>.
- Aghamohammadpour, Ali, Ebrahim Mahdipour, and Iman Attarzadeh. "Architecting Threat Hunting System Based on the DODAF Framework." *The Journal of Supercomputing* 79, no. 4 (September 23, 2022): 4215-4242. doi:10.1007/s11227-022-04808-6. <https://doi.org/10.1007/s11227-022-04808-6>.
- Ahmed, Mohiuddin, Sheikh Rabiul Islam, Adnan Anwar, Nour Moustafa, and Al-Sakib Khan Pathan. "Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence ." *Studies in Computational Intelligence* 1025, (2022): 280. <https://doi.org/10.1007/978-3-030-96630-0>.
- Akçay, Tolga. *The AI Compass - Welcome to the World of Artificial Intelligence*. United States: 2021.
- Alabdulatif, Abdullah and Navod Neranjan Thilakarathne. *Bio-Inspired Internet of Things: Current Status, Benefits, Challenges, and Future Directions*. Vol. 8 MDPI AG, 2023.
- Alake Richmond. "Loss Functions in Machine Learning Explained." Accessed March 7, 2024. <https://www.datacamp.com/tutorial/loss-function-in-machine-learning>.
- Al-Dosari, Khalifa, Noora Fetais, and Murat Kucukvar. "Artificial Intelligence and Cyber Defense System for Banking Industry: A Qualitative Study of AI Applications and Challenges." *Cybernetics and Systems* 55, no. 2 (Aug 12, 2022): 1-29. doi:10.1080/01969722.2022.2112539. <https://doi.org/10.1080/01969722.2022.2112539>.
- Alshaibi, Ahmed, Mustafa Al-Ani, Abeer Al-Azzawi, Anton Konev, and Alexander Shelupanov. "The Comparison of Cybersecurity Datasets." *Data* 7, no. 2 (January 29, 2022): 1-19. doi:10.3390/data7020022. <https://doi.org/10.3390/data7020022>.
- Amazon Web Services. "A Realistic Cyber Defense Dataset (CSE-CIC-IDS2018)." Accessed May 4, 2024. <https://registry.opendata.aws/cse-cic-ids2018/>.
- Amazon Web Services. "What is Artificial Intelligence (AI)?" Accessed April 24, 2024. <https://aws.amazon.com/what-is/artificial-intelligence/>.
- Amazon Web Services. "What is Hyperparameters Tuning?" Accessed March 26, 2024. <https://aws.amazon.com/what-is/hyperparameter-tuning/>.
- Apruzzese, Giovanni, Pavel Laskov, Edgardo Montes De Oca, Wissam Mallouli, Luis Brdalo Rapa, Athanasios Vasileios Grammatopoulos, and Fabio Di Franco. "The Role of Machine Learning in Cybersecurity." *Digital Threats: Research and Practice* 4, no. 1 (March 7, 2023): 1-38. doi:10.1145/3545574. <http://arxiv.org/abs/2206.09707>.
- Araya, Daniel. *Artificial Intelligence for Defence and Security - Special Report*. Waterloo: 2022.

- Asemi, Asefeh, Andrea Ko, and Mohsen Nowkarizi. "Intelligent Libraries: A Review on Expert Systems, Artificial Intelligence, and Robot." *Library Hi Tech* 39, no. 2 (June 6, 2020): 412-434. doi:10.1108/lht-02-2020-0038. <https://doi.org/10.1108/LHT-02-2020-0038>.
- Bengesi, Staphord, Hoda El-Sayed, Kamruzzaman Sarker, Yao Houkpati, John Irungu, and Timothy Oladunni. *Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers*. United States, Ithaca: Cornell University Library arXiv.org, 2022.
- Brownlee, Jason. "What is the Difference between Test and Validation Datasets?" Accessed March 27, 2024. <https://machinelearningmastery.com/difference-test-validation-datasets/>.
- Burt, Andrew. "How to Red Team a Gen AI Model." Accessed March 14, 2024. <https://hbr.org/2024/01/how-to-red-team-a-gen-ai-model>.
- Canadian Armed Forces. *Canadian Armed Forces Digital Campaign Plan* His Majesty the King in Right of Canada, as represented by the Minister of Environment and Climate Change, 2022.
- Canadian Centre for Cyber Security. *Artificial Intelligence - ITSAP.00.040* Communications Security Establishment, 2023.
- Canadian Centre for Cyber Security. *Generative Artificial Intelligence - ITSAP.00.041* Communications Security Establishment, July. <https://www.cyber.gc.ca/en/guidance/generative-artificial-intelligence-ai-itsap00041>.
- Canadian Institute for Cybersecurity, (CIC). "Datasets." Accessed March 6, 2024. <https://www.unb.ca/cic/datasets/index.html>.
- Chakraborty, Utpal, Soumyadeep Roy, and Sumit Kumar. *Rise of Generative AI and Chat GPT - Understand how Generative AI and ChatGPT are Transforming and Reshaping the Business World*. London: BPB Publications, 2023.
- Communications Security Establishment. *Canadian Centre for Cyber Security - National Cyber Threat Assessment 2023-2024* His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2022.
- Czech, Johannes. "Distributed Methods for Reinforcement Learning Survey." In *Reinforcement Learning Algorithms: Analysis and Applications*, edited by Belousove, Boris, Hany Abdulsamad, Pascal Klink, Simone Parisi and Jan Peters, 151-161: Springer Nature Switzerland, 2021.
- Das, Rammanohar and Raghav Sandhane. "Artificial Intelligence in Cyber Security." *Journal of Physics: Conference Series* 1964, no. 4 (Jul 23, 2021): 1-10. doi:10.1088/1742-6596/1964/4/042072. <https://doi.org/10.1088/1742-6596/1964/4/042072>.
- Databricks. "Dataset." Accessed April 2, 2024. <https://www.databricks.com/glossary/what-is-dataset>.

- De, Moloy. "Hidden Layers of A Neural Network." Accessed March 22, 2024. <https://community.ibm.com/community/user/ai-datascience/blogs/moloy-de1/2023/04/16/hidden-layers-of-a-neural-network>.
- Defence Stories. "DM/CDS Message: Launch of the DND/CAF Artificial Intelligence Strategy." Accessed April 15, 2024. <https://www.canada.ca/en/department-national-defence/maple-leaf/defence/2024/03/dm-cds-message-launch-dnd-caf-artificia-intelligence-strategy.html>.
- Dwivedi, Rudresh, Devam Dave, Het Naik, Smriti Singhal, Rana Omer, Pankesh Patel, Bin Qian, et al. "Explainable AI (XAI): Core Ideas, Techniques, and Solutions." *ACM Computing Surveys* 55, no. 9 (January, 2023). doi:10.1145/3561048. <https://dl.acm.org/doi/10.1145/3561048>.
- Ebert, Christof and Maximilian Beck. "Artificial Intelligence for Cybersecurity." *IEEE Software* 40, no. 6 (Nov, 2023). doi:10.1109/ms.2023.3305726. <https://ieeexplore-ieee.org.cfc.idm.oclc.org/document/10339105>.
- Fadel, Soufiane. "Explainable Machine Learning, Game Theory, and Shapley Values: A Technical Review." Accessed March 23, 2024. <https://www.statcan.gc.ca/en/data-science/network/explainable-learning>.
- Gartner. "Life at Gartner." Accessed May 4, 2024. <https://www.gartner.com/en>.
- Google. "Data Prep - Construct Your Dataset." Accessed April 3, 2024. <https://developers.google.com/machine-learning/data-prep/construct/construct-intro>.
- Google. "Data Prep - Imbalanced Data." Accessed April 13, 2024. <https://developers.google.com/machine-learning/data-prep/construct/sampling-splitting/imbalanced-data>.
- Google. "Machine Learning Crash Course - Descending into ML." Accessed April 2, 2024. <https://developers.google.com/machine-learning/crash-course/descending-into-ml/training-and-loss>.
- Google. "Machine Learning Crash Course - Framing ." Accessed April 2, 2024. <https://developers.google.com/machine-learning/crash-course/framing/ml-terminology>.
- Google. "Machine Learning Crash Course - Reducing Loss." Accessed April 23, 2024. <https://developers.google.com/machine-learning/crash-course/reducing-loss/an-iterative-approach>.
- Google. "Machine Learning Crash Course - Validation Set: Another Partition." Accessed April 19, 2024. <https://developers.google.com/machine-learning/crash-course/validation/another-partition>.
- Google. "Machine Learning Glossary - Parameter." Accessed April 24, 2024. <https://developers.google.com/machine-learning/glossary#parameter>.
- Google. "Machine Learning Glossary - Training Set." Accessed April 3, 2024. [https://developers.google.com/machine-learning/glossary#training\\_set](https://developers.google.com/machine-learning/glossary#training_set).

- Google. "What are AI Hallucinations?" Accessed March 18, 2024. <https://cloud.google.com/discover/what-are-ai-hallucinations>.
- Google. "What is Machine Learning?" Accessed April 2, 2024. <https://developers.google.com/machine-learning/intro-to-ml/what-is-ml>.
- Government of Canada. "Algorithmic Impact Assessment Tool." Accessed March 15, 2024. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/algorithmic-impact-assessment.html>.
- Government of Canada. "The Artificial Intelligence and Data Act (AIDA) – Companion Document." Accessed March 16, 2024. <https://ised-isde.canada.ca/site/innovation-better-canada/en/artificial-intelligence-and-data-act-aida-companion-document>.
- Government of Canada. "Directive on Automated Decision-Making." Accessed January 12, 2024. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592>.
- Government of Canada. "Directive on Service and Digital." Accessed March 17, 2024. <https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32601>.
- Government of Canada. "Gender-Based Analysis Plus (GBA Plus)." Accessed April 22, 2024. <https://www.canada.ca/en/women-gender-equality/gender-based-analysis-plus.html>.
- Government of Canada. "Guide on the use of Generative AI." Accessed March 21, 2024. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html>.
- Government of Canada. "Pan-Canadian Artificial Intelligence Strategy." Accessed March 16, 2024. <https://ised-isde.canada.ca/site/ai-strategy/en>.
- Government of Canada. "Responsible use of Artificial Intelligence (AI)." Accessed March 17, 2024. <https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai.html>.
- Government of Canada. "Voluntary Code of Conduct on the Responsible Development and Management of Advanced Generative AI Systems." Accessed March 17, 2024. <https://ised-isde.canada.ca/site/ised/en/voluntary-code-conduct-responsible-development-and-management-advanced-generative-ai-systems>.
- Gunning, David and David W. Aha. "DARPA's Explainable Artificial Intelligence Program." *AI Magazine* 40, no. 2 (2019): 44-58. <https://login.cfc.idm.oclc.org/login?qurl=https%3A%2F%2Fwww.proquest.com%2Fscholarly-journals%2Fdarpas-explainable-artificial-intelligence%2Fdocview%2F2258093718%2Fse-2%3Faccountid%3D9867>.
- Gupta, Sparsh. "The 7 most Common Machine Learning Loss Functions ." Accessed March 7, 2024. <https://builtin.com/machine-learning/common-loss-functions>.
- Haenlein, Michael and Andreas Kaplan. "A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence." *California Management Review* 61, no. 4

- (July 17, 2019).  
doi:10.1177/0008125619864925. <https://doi.org/10.1177/0008125619864925>.
- IBM. "What are AI Hallucinations?" Accessed March 18, 2024. <https://www.ibm.com/topics/ai-hallucinations>.
- IBM. "What is an AI Model?" Accessed Dec 20, 2023. <https://www.ibm.com/topics/ai-model>.
- IBM. "What is an Attack Surface?" Accessed March 14, 2024. <https://www.ibm.com/topics/attack-surface>.
- Janiesch, Christian, Patrick Zschech, and Kai Heinrich. "Machine Learning and Deep Learning." *Electronic Markets* 31, no. 3 (April 8, 2021): 685-695. doi:10.1007/s12525-021-00475-2. <https://doi-org.cfc.idm.oclc.org/10.1007/s12525-021-00475-2>.
- Khoulimi, Hind, Mohamed Lahby, and Othman Benammar. "Chapter 2 an Overview of Explainable Artificial Intelligence for Cyber Security." In *Explainable Artificial Intelligence for Cyber Security - Next Generation Artificial Intelligence*. Vol. 1025, 280. Warsaw: Springer, 2022.
- Kumar, Pradeep and B. R. Prasad Babu. "Investigating Open Issues in Swarm Intelligence for Mitigating Security Threats in MANET." *International Journal of Electrical & Computer Engineering* 5, no. 5 (October, 2015): 1194-1201. <http://iaesjournal.com/online/index.php/IJECE/issue/archive>.
- Liu, Longhui, Yang Zhou, Qing Xu, Qunshan Shi, and Xiaofei Hu. "Improved Technique for Order of Preference by Similarity to Ideal Solution Method for Identifying Key Terrain in Cyberspace Asset Layer." *Plos One* 18, no. 7 (July 13, 2023): 1-20. doi:10.1371/journal.pone.0288293. 288293. <https://doi.org/10.1371/journal.pone.0288293>
- Mehta, Mayuri, Vasile Palade, and Indranath Chatterjee. "Explainable AI - Foundations, Methodologies and Applications." *Springer* 232, (2023): 273. doi:10.1007/978-3-031-12807-3. <https://doi.org/10.1007/978-3-031-12807-3>.
- Mohamed, Nachaat. "Current Trends in AI and ML for Cybersecurity: A State-of-the-Art Survey." *Cogent Engineering* 10, no. 2 (October 25, 2023): 1-30. doi:10.1080/23311916.2023.2272358. <https://doi.org/10.1080/23311916.2023.2272358>.
- National Defence. *B-GL-300-001-FP-001 — Land Operations* Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2008.
- National Defence. *Canadian Armed Forces Joint Doctrine Note - 2017-02 - Cyber Operations* Her Majesty the Queen as represented by the Minister of National Defence, 2107.
- National Defence. *The Department of National Defence and Canadian Armed Forces - Data Strategy*. Ottawa, Ont: Her Majesty the Queen as represented by the Minister of National Defence, 2019.

- National Defence. *The Department of National Defence and Canadian Armed Forces Artificial Intelligence Strategy*. Ottawa: His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2024a.
- National Defence. "Our North, Strong and Free: A Renewed Vision for Canada's Defence." . Accessed April 15, 2024. <https://www.canada.ca/en/department-national-defence/news/2024/04/our-north-strong-and-free-a-renewed-vision-for-canadas-defence.html>.
- National Defence. *Our North, Strong and Free: A Renewed Vision for Canada's Defence* His Majesty the King Right of Canada as represented by the Minister of National Defence, 2024c.
- National Defence. *Pan-Domain Force Employment Concept - Prevailing in a Dangerous World* His Majesty the King in Right of Canada, as represented by the Minister of National Defence, 2023.
- National Defence. *Strong Secure Engaged - Canada's Defence Policy* Her Majesty the Queen in Right of Canada, as represented by the Minister of National Defence, 2017.
- Nelson, Daniel. "Generative Vs. Discriminative Machine Learning Models." Accessed February 28, 2024. <https://www.unite.ai/generative-vs-discriminative-machine-learning-models/>.
- Nelson, Daniel. "What is a Generative Adversarial Network (GAN)?" Accessed March 26, 2024. <https://www.unite.ai/what-is-a-generative-adversarial-network-gan/>.
- Nelson, Daniel. "What is a KNN (K-Nearest Neighbors)?" Accessed May 6, 2024. <https://www.unite.ai/what-is-k-nearest-neighbors/>.
- Nelson, Daniel. "What is an Autoencoder?" Accessed March 22, 2024. <https://www.unite.ai/what-is-an-autoencoder/>.
- Nelson, Daniel. "What is K-Means Clustering?" Accessed May 6, 2024. <https://www.unite.ai/what-is-k-means-clustering>.
- Pfob, André, Sheng-Chieh Lu, and Chris Sidey-Gibbons. "Machine Learning in Medicine: A Practical Introduction to Techniques for Data Pre-Processing, Hyperparameter Tuning, and Model Comparison." *BMC Medical Research Methodology* 22, no. 1 (-11-01, 2022): 1-15. doi:10.1186/s12874-022-01758-8. <https://doi.org/10.1186/s12874-022-01758-8>.
- Phillips, P. Jonathon, Carina A. Hahn, Peter C. Fontana, Amy N. Yates, Kristen Greene, David A. Broniatowski, and Mark A. Przybocki. "Four Principles of Explainable Artificial Intelligence." *NIST Interagency/Internal Report (NISTIR) 8312*, (September 29, 2021). doi:10.6028/nist.ir.8312. <https://doi.org/10.6028/NIST.IR.8312>.
- Raja Sindiramutty, Siva. "Autonomous Threat Hunting: A Future Paradigm for AI-Driven Threat Intelligence." Taylors University Subang Jaya Malaysia, 2023.
- Schmitt, Markus. "Machine Learning Architecture: The Code Components." Accessed March 28, 2024. <https://www.datarevenue.com/en-blog/machine-learning-project-architecture>.

- Shaukat, Kamran, Suhui Luo, Vijay Varadharajan, Ibrahim A. Hameed, Shan Chen, Dongxi Liu, and Jiaming Li. "Performance Comparison and Current Challenges of using Machine Learning Techniques in Cybersecurity." *Energies* 13, no. 10 (May 15, 2020): 1-27. doi:10.3390/en13102509. <https://doi.org/10.3390/en13102509>.
- Siva Kumar, Ram Shankar, Magnus Nystrom, John Lambert, Andrew Marshall, Mario Goertzel, Andi Comissoneru, Matt Swann, and Sharon Xia. "Adversarial Machine Learning-Industry Perspectives." *2020 IEEE Security and Privacy Workshops (SPW)* (May, 2020): 69-75. doi:10.1109/spw50608.2020.00028. <https://ieeexplore-ieee-org.cfc.idm.oclc.org/document/9283867>.
- splunk. *The PEAK Threat Hunting Framework - Modernized Hunting for the Evolving Threat Landscape*. United States: Splunk Inc., 2023.
- Stevens, Tim. "BMW showed Off Hallucination-Free AI at CES 2024." . Accessed March 20, 2024. <https://arstechnica.com/cars/2024/01/bmws-ai-powered-voice-assistant-at-ces-2024-sticks-to-the-facts/>.
- Storey, Veda C., Roman Lukyanenko, Wolfgang Maass, and Jeffrey Parsons. "Viewpoint - Explainable AI." *Communications of the ACM* 65, no. 4 (April, 2022): 27-29. doi:10.1145/3490699. <https://doi.org/10.1371/journal.pone.0231166>.
- Taddeo, Mariarosaria, Tom Mccutcheon, and Luciano Floridi. "Trusting Artificial Intelligence in Cybersecurity is a Double-Edged Sword." *Nature Machine Intelligence* 1, no. 12 (November 11, 2019): 557-560. doi:10.1038/s42256-019-0109-1. <https://doi.org/10.1038/s42256-019-0109-1>.
- The MITRE Corporation. "Mitre Att&ck®." Accessed March 14, 2024. <https://attack.mitre.org/>.
- Truong, Thanh Cong, Quoc Bao Diep, and Ivan Zelinka. "Artificial Intelligence in the Cyber Domain: Offense and Defense." *Symmetry* 12, no. 3 (Mar 01, 2020): 1-24. doi:10.3390/sym12030410. <https://doi.org/10.3390/sym12030410>.
- UK National Cyber Security Centre (GCHQ). *Guidelines for Secure AI System Development*. Washington, D.C: UK NCSC, US Cybersecurity and Infrastructure Security Agency (CISA), 2023.
- Van De Sande, Davy, Michel E. Van Genderen, Jim M. Smit, Joost Huiskens, Jacob J. Visser, Robert E. R. Veen, Edwin Van Unen, Oliver Hilgers Ba, Diederik Gommers, and Jasper Van Bommel. "Developing, Implementing and Governing Artificial Intelligence in Medicine: A Step-by-Step Approach to Prevent an Artificial Intelligence Winter." *BMJ Health Care Inform* 29, no. 1 (-02, 2022). doi:10.1136/bmjhci-2021-100495. <https://doi.org/10.1136/bmjhci-2021-100495>.
- Ventre, Daniel. *Artificial Intelligence, Cybersecurity and Cyber Defence*. Hoboken: John Wiley & Sons, Incorporated, 2020.
- Yao, Kaiqi, Qianzhen Zhang, Bin Liu, and Xianqiang Zhu. "Cyberspace Knowledge Base for Key Terrain Identification." *Electrical and Electronics Engineers (IEEE)* (August 21, 2023): 728-735. doi:10.1109/AINIT59027.2023.10212951. <https://ieeexplore-ieee-org.cfc.idm.oclc.org/document/10212951>.

- Zhang, Longbin, Zhijun Li, Yingbai Hu, Christian Smith, Elena M. Gutierrez Farewik, and Ruoli Wang. *Ankle Joint Torque Estimation using an EMG-Driven Neuromusculoskeletal Model and an Artificial Neural Network Model*. Vol. 18 Institute of Electrical and Electronics Engineers (IEEE), 2021.
- Zhang, Zhimin, Huansheng Ning, Feifei Shi, Fadi Farha, Yang Xu, Jiabo Xu, Fan Zhang, and Kim-Kwang Raymond Choo. "Artificial Intelligence in Cyber Security: Research Advances, Challenges, and Opportunities." *Artificial Intelligence Review* 55, no. 2 (March 13, 2021). doi:10.1007/s10462-021-09976-0. <https://doi.org/10.1007/s10462-021-09976-0>.