Hiring a Hackable Colleague

The Dangers of Adversarial AI in an Automated Intelligence Cycle

Bo Lievisse Adriaanse

University of Glasgow ID: 2781365
Dublin City University ID: 22111310
Charles University ID: 9982524

Presented in partial fulfilment of the requirements for the Degree of International Master in Security, Intelligence and Strategic Studies

Word Count: 24.200
Supervisor: Dr. Rhys Crilley
Date of Submission: 27/06/2024

Abstract

The advent of Artificial Intelligence (AI) has opened up new doors for intelligence analysis, promising unprecedented speed and accuracy. Indeed, as AI continues to evolve, its contribution to how intelligence is gathered, processed, and interpreted becomes increasingly apparent. However, employing this cutting-edge technology requires careful consideration of its potential limitations and its impact on the viability of the intelligence cycle in its traditional form. This dissertation explores the intricate intersection between AI and the intelligence cycle, highlighting the threats posed by adversarial AI – techniques designed to deceive or manipulate AI systems. By dissecting key stages of the intelligence cycle—data collection, preprocessing, and analysis—this dissertation underlines how advanced AI systems can be exploited to effectuate misleading or harmful results. Several cases are presented, studying Convolutional Neural Network (CNN)-powered Unmanned Aerial Vehicles (UAVs), Natural Language Processing (NLP) systems, and Generative AI (GenAI). These case studies illustrate how different adversarial attacks, such as evasion, poisoning, and Dual Denial of Decision (DDoD) attacks, can have real-world implications on the accuracy and viability of the traditional intelligence cycle. The findings emphasise the critical need for a balanced approach that leverages the benefits of AI while fortifying defence mechanisms against adversarial threats. To address these challenges and accurately account for AI's revolutionary impact on intelligence, this dissertation presents the Augmented Intelligence Cycle, a reimagined version of the traditional intelligence cycle. This new model emphasises continuous collaboration between humans and AI to ensure the secure and effective integration of AI in the domain of intelligence.

Keywords: Artificial Intelligence | Adversarial AI | Intelligence Cycle | Intelligence Automation

Preface

"Success in creating AI would be the biggest event in human history. Unfortunately, it might also be the last, unless we learn how to avoid the risks." – Stephen Hawking

Humans tend to experience a strong sense of nostalgia when reflecting on the past. Everything seems to have been better yesterday, and even more so the day before yesterday – A fallacious thought, as we tend to forget the bad more quickly than the good. Don't get me wrong, my brain works the exact same way.

While Hawking's warning is valid (the relevance of this dissertation hinges on it), it also perpetuates the idea that a future that is uncertain is inherently negative. Similarly, the discourse surrounding AI seems to be dominated by doom scenarios, whether it be in news articles, on social media, or in science fiction. Few of these narratives explore AI's potential to benefit humanity.

I wrote this dissertation with the best intentions, convinced that raising awareness about the limitations of AI is crucial for steering its responsible development and use. AI is not just a gadget for generating amusing images; it is a double-edged sword that has already cut deeply with profound and unjust consequences. While I still stand by this message, it should come with an important sidenote: If we want to avoid our dystopian fears from turning into a self-fulfilling prophecy, we must balance our focus on risks with stories of innovation and hope.

Humans created AI, and humans will decide how to use it. While understanding the risks is imperative, we should therefore also cultivate a sense of future nostalgia—envisioning a positive future shaped by AI. This dissertation has inspired me to do just that.

Beyond the scope of this work, I want to thank those who have supported me throughout this journey and contributed to the completion of this dissertation.

To my supervisor, *Dr. Rhys Crilley*, thank you for your continuous guidance, constructive feedback, reassurance. Your insights and encouragement have been instrumental in transforming my ideas into a dissertation of academic merit.

To my colleagues at *TNO*, thank you for deepening my understanding of adversarial AI and related fields. Our many discussions and talks have inspired me and have further fuelled my enthusiasm for this topic.

To my *friends, family, and Rick*, thank you for your unwavering support during this intense period and for always bringing a smile to my face. My success is yours, and I hope I have made you proud.

– Bo Lievisse Adriaanse

List of Abbreviations

AAI = Adversarial Artificial Intelligence

ACINT = Acoustical Intelligence

AI = Artificial Intelligence

AP = Average Precision

AdvTexture = Adversarial Texture

CNN = Convolutional Neural Network

COMINT = Communications Intelligence

CV = Computer Vision

CYBINT = Cyber Intelligence

DDoD = Dual Denial of Decision

DE_DAC = Differential Evolution-based Dual Adversarial Camouflage

DL = Deep Learning

DNN = Deep Neural Network

ELECTRO-OPTINT = Electro-optical Intelligence

ELINT = Electronic Signals Intelligence

FCA = Full-Coverage Camouflage Attack

FISINT = Foreign Instrumentation Signals Intelligence

GEOINT = Geospatial Intelligence

GenAI = Generative AI

Grad-CAM = Gradient-Weighted Class Activation Maps

HUMINT = Human Intelligence

IMINT = Imagery Intelligence

IRINT = Infrared Intelligence

LLM = Large Language Model

MAS = Multi-Agent System

MASINT = Measurement and Signature Intelligence

ML = Machine Learning

NAP = Naturalistic-Adversarial-Patch

NE = Named Entity

NER = Named Entity Recognition

NLP = Natural Language Processing

NUCINT = Nuclear Intelligence

OSINT = Open Source Intelligence

RADINT = Radar Intelligence

RISOPA = Rapid Imperceptible Strong One-Pixel Attacks

SIGINT = Signals Intelligence

SOCMINT = Social Media Intelligence

TC-EGA = Toroidal-Cropping-based Expandable Generative Attack

UAV = Unmanned Aerial Vehicle

XAI = Explainable Artificial Intelligence

Contents

Chapter 1

Introduction

The rise of Artificial Intelligence (AI) marks one of the most significant technological advancements of our time. From transforming industries to revolutionising daily life, AI's capabilities in data processing, pattern recognition, and decision-making show tremendous promise (Makridakis, 2017; Erskine & Miller, 2024). This revolutionary technology has understandably become a critical determinant of international geopolitical status, shaping both offensive and defensive capabilities (Chow & Perrigo, 2023; King, 2024). Indeed, beyond generating funny images, AI has quickly made its way to the battleground. As noted by King "AI has already begun to play a significant role in military operations and is likely to be more important in the future" (2024, p. 1).

This is exemplified by the Israel-Palestine conflict, where the IDF employs AI-powered systems such as Lavender and "Where's Daddy?" to automatically identify and target suspects – a concerning development, as these systems have been reported to misidentify large numbers of citizens as combatants (Abraham, 2024). This problematic use of AI in high-stakes environments underscores the question of whether governments are ready and able to integrate AI thoughtfully and securely. Meanwhile, AI's popularity surges, with governments and private corporations making serious investments, and companies like Nvidia, TSMC, and ASML growing from lucrative businesses into geopolitical pawns (Goldman Sachs, 2023).

Indeed, with the traditional battlefield expanding into the digital domain, AI has become a crucial geopolitical asset (Wen & Song, 2018). Yet, in an increasingly tense and geopolitically shifting world, the urgency of AI implementation seems to outpace a comprehensive understanding of its risks. Importantly, AI is not a magical solution but a technical tool with significant limitations and vulnerabilities. This is echoed by Erskine and Miller (2024) and Logan (2024), who stress that AI will either support intelligence endeavours or lead to critical errors in high-stakes scenarios. Echoing this is a substantial body of literature on Adversarial AI (AAI), detailing various attack methods to manipulate AI systems (Amoroso & Tamburrini, 2020; Biggio & Roli, 2018; Christie et al., 2023; Goodfellow et al., 2018).

However, as King (2024) stresses, the literature's focus on the implications of AI and AAI on applications such as autonomous weaponry is disproportionately large compared to its real-world implementations. In practice, AI is more frequently used for intelligence purposes instead, specifically to automate parts of the intelligence cycle, a key framework in the intelligence domain (Blanchard & Taddeo, 2023; King, 2024; Mitchell et al., 2019). Problematically, despite the rapid increase of AI in the intelligence domain, academic research on this topic does not reflect its real-world significance, with even less attention given to AAI.

Thus, although research on the intelligence cycle, AI, and AAI exists, these areas of study rarely intersect, underscoring the critical need to assemble these pieces cohesively. Moreover, AI's impact on security and intelligence is so big that it requires us to rethink whether the intelligence cycle is still usable in its current form. As such, it becomes imperative to identify where the intelligence cycle needs adjustment to effectively integrate AI.

To address this research gap, this dissertation aims to answer the following research question:

> **How can the intelligence cycle be adapted to ensure the effective and secure integration of AI technologies, considering their increased use and vulnerability to adversarial manipulation?**

This question addresses the dual challenges of enhancing the intelligence cycle with AI while mitigating the associated risks. By focusing on this inquiry, the dissertation aims to provide actionable insights for intelligence agencies on adjusting their processes and frameworks to safely and effectively accommodate the integration of AI.

By systematically integrating these puzzle pieces, this dissertation offers a significant and novel contribution to both academia and practical fields. It underscores the necessity of a holistic understanding of AI's impact on intelligence operations, advocating for a balance between rapid implementation and thorough risk assessment. This approach culminates in the creation of a revised intelligence cycle that accounts for these transformative changes. By doing so, the dissertation not only advances theoretical knowledge but also provides practical insights, ensuring that intelligence operations remain effective and resilient in the face of AI-driven advancements.

The methodology utilises an instrumental case study approach, chosen for its ability to provide insights into broader phenomena through the examination of individual cases (Stake, 2005). The cases were selected based on stages of the intelligence cycle and their potential for automation as well as manipulation. The intelligence cycle serves as a guiding framework due to its comprehensive structure and long-standing application in intelligence operations (Kent, 1966; Lowenthal, 2020). The intelligence cycle consists of five steps: planning and direction, collection, processing, analysis, and dissemination. However, for this study, only the collection, processing, and analysis phases were selected, as these appear most suitable for automation and form the core of the final intelligence product (Jensen et al., 2022). Each steps was systematically analysed to identify potential automation using AI technologies and corresponding vulnerabilities, allowing for an evaluation of the cycle's effectiveness when partially automated. At the same time, the diverse cases provide a comprehensive understanding of the interconnected nature of the cycle's subsequent stages, offering a detailed assessment of the traditional cycle's limitations in the context of AI integration.

The data utilised in the individual cases consists of existing literature, which is specifically analysed using the principles of an integrative literature review. This approach allows for the synthesis of existing research while addressing the dynamic and ever-evolving nature of AI, AAI, and the intelligence cycle (Lubbe et al., 2020; Torraco, 2016). Researching intelligence is inherently challenging due to national security concerns, resulting in much of the relevant data being classified (Kerr, 2023). Furthermore, while it is not definitively known, the implementation of AI in intelligence analysis appears to be in its early stages, particularly since AI has only recently achieved a level of performance that justifies its integration into high-stakes national security matters (Intel AI, 2019). Still, the rapid pace of development underscores the urgent need to identify vulnerabilities before AI is fully implemented, if it is not already. Therefore, despite the inherent challenges of researching national security topics, this dissertation focuses on publicly available data, ensuring a comprehensive yet ethical examination of the subject. Specifically, by focusing on state-of-the-art literature on AI and AAI, and combining these insights with the intelligence cycle framework, this research addresses a critical research gap in a dynamic and ever-evolving field.

The primary aim of this dissertation is to analyse the implications of AI in order to rethink and redefine the intelligence cycle in this age of AI. To achieve this, the dissertation is structured as follows: First, a thorough review of existing literature on AI in intelligence analysis, AAI, and the traditional intelligence cycle is conducted. Following this, the methodology section details the research design, including the selection criteria for case studies and data collection methods. The subsequent results section presents case studies that examine real-world instances of AI integration in intelligence analysis, illustrating the realistic threats of AAI to intelligence cycle automation. A comprehensive discussion of these results follows in which the complexities of automating the intelligence cycle are highlighted. Based on these insights, I propose an augmented intelligence cycle to effectively accommodate human-AI collaboration. The dissertation concludes with a summary of key findings and recommendations for intelligence agencies on the secure and effective implementation of AI, as well as the dissertation's limitations and suggestions for future research.

The dissertation concludes that although AI offers substantial potential to enhance various stages of the intelligence cycle, it also introduces new vulnerabilities that adversaries may exploit. To effectively integrate AI, a balanced approach is necessary, where human agents and AI-enabled systems work in synergy while mitigating adversarial threats. The case studies underscore the importance of context-specific strategies and the need to revise the intelligence cycle, transforming it into an augmented cycle that supports responsible and secure AI integration within the intelligence domain.

Chapter 2

Literature Review

2.1. Defining Intelligence

The concept of intelligence is a complex and multifaceted construct that has been the subject of much scholarly inquiry and debate over the years. Despite numerous attempts to define and conceptualise intelligence, a definitive, all-encompassing definition remains elusive. As Breakspear notes, "Intelligence is widely misunderstood" (2013, p. 1). Kahn proceeds to define intelligence in the most general sense as *information* (2001, p. 79). However, this simplistic view fails to account for the intricate cognitive processes and organisational aspects underlying the concept of intelligence. Hedley provides a slightly more nuanced definition, arguing that intelligence is not just information, but information plus insight (p. 2007, 212). While this definition is an improvement, it still falls short of capturing the full complexity of the field.

Lowenthal (2020) offers a more comprehensive understanding of intelligence that considers the multifaceted nature of the concept. In his view, intelligence is

> The process by which specific types of information important to national security are requested, collected, analyzed, and provided to policy makers; the products of that process; the safeguarding of these processes and this information by counterintelligence activities; and the carrying out of operations as requested by lawful authorities" (Lowenthal, 2020, p.16).

This definition emphasises that intelligence is not simply a matter of gathering information, but rather a multifaceted process that requires careful planning, execution, and oversight. By emphasising intelligence as both a product, process and organisational endeavour, his definition offers a more nuanced view of intelligence that considers the complex cognitive processes and organizational aspects involved. This framework, particularly focusing on intelligence as a process, will serve as the foundation for the dissertation, providing a solid basis for further exploration and analysis.

2.2. Understanding The Intelligence Cycle

To gain a better understanding of intelligence as a process, it is important to explore its constituent parts and practices. Specifically, this involves exploring the nuances of requesting, collecting, analysing, and providing information that is necessary for the creation of intelligence. The intelligence cycle stands as a renowned model encapsulating this intricate process. It has evolved into the quintessential framework for comprehending intelligence operations, endorsed by prominent figures in the field such as Kent (1966), Lowenthal (2020), and Clark (2022). Intelligence agencies worldwide have embraced this model or variations thereof. For example, while the UK intelligence cycle comprises only four steps, the most widely accepted rendition, illustrated in Figure 1, delineates it into five distinct stages: 1) Planning and Direction, 2) Collection, 3) Processing, 4) Analysis and Production, and 5) Dissemination (Phythian, 2013).

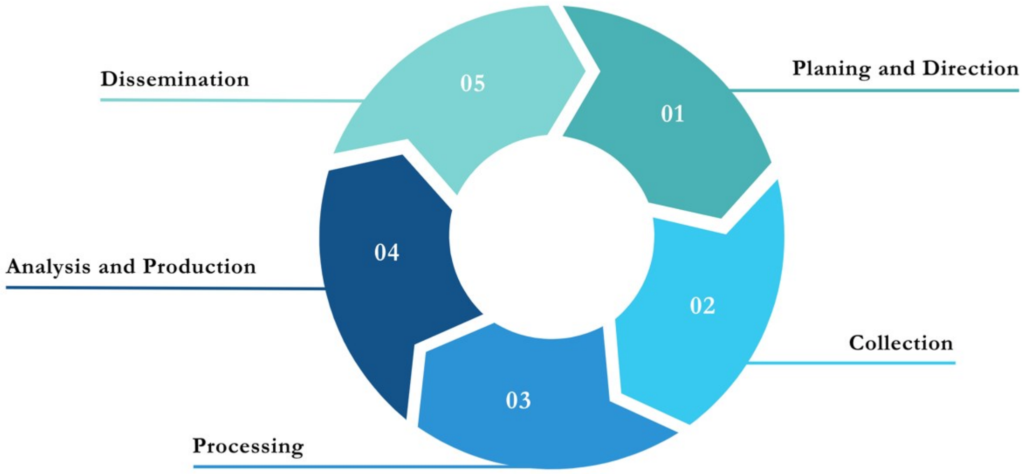


Figure 1: The five stages of the Intelligence Cycle

During the initial phase of the intelligence cycle, referred to as Planning and Direction, stakeholders and policymakers establish their specific intelligence requirements and priorities regarding a particular target, situation, or issue (Phythian, 2013). This stage establishes the strategic framework that guides subsequent intelligence activities. After receiving the client's requirements, the process of collecting data begins. This involves obtaining raw data from various sources, which are classified into different categories called "INTs", standing for intelligence (Phythian, 2013). Table 1 provides a summary and explanation of these intelligence categories based on Henrico and Putter's (2024) thorough systematic review of the various intelligence collection disciplines. As technology advances, the amount of available data not only increases but also changes in nature. To keep up with this complexity, new techniques for data analysis and interpretation are continuously being developed, ultimately leading to the emergence of new categories based on existing ones (Henrico & Putter, 2024).

OSINT is a good example of a rapidly developing intelligence collection method driven by new technologies such as social media and AI (Evangelista et al., 2020). This approach may intersect with other disciplines, for example overlapping with social media intelligence (SOCMINT) when analysing Instagram content or converging with geospatial intelligence (GEOINT) when leveraging features like Google Maps' satellite imagery. As this discipline allows valuable data to be obtained faster and at a lower cost, this method is becoming increasingly popular (Pastor-Galindo et al., 2020). Understandably, the ascendance and waning of different intelligence disciplines are thus largely propelled by the latest technological innovations. Notably, there seems to be one exception: human intelligence (HUMINT), which hinges on human interaction and therefore diverges from the reliance on technology seen in other disciplines.

While the information gathered through these methods is crucial, it is not intelligence yet. To be transformed into intelligence, raw information must pass through the processing and analysis stages. The processing stage is predominantly technical and involves filtering and preparing information for analysis through methods like decryption, translation, and condensation (Phythian, 2013). Analysis, the subsequent step, is when the data itself is turned from information into intelligence. This entails synthesizing, evaluating, and dissecting the, often fragmented or conflicting, data into a final intelligence product (Jensen et al., 2022). The analysts responsible for this particular stage of the process meticulously examine the relevance, credibility, and accuracy of the gathered information. They then proceed to integrate it into a comprehensive narrative, providing context and offering assessments and interpretations of its significance within the given framework (Pherson & Heuer, 2015). This part of the process is commonly known as analytical tradecraft, and understandably so, given the significant skill required to produce high-quality intelligence outputs (Pherson & Heuer, 2015). Hence, despite gathering sufficient and relevant data, incorrect interpretation essentially renders that information useless.

Omand (2023) notes how, in a general sense, intelligence analysis can support a decision-maker in four distinct ways; by providing 1) situational awareness, 2) explanations, 3) predictions or estimates, and 4) warnings. While the methods used to achieve and present these objectives have significantly changed over time, these purposes have remained unchanged. Intelligence analysis is far from an exact science and its accuracy can only be assessed in retrospect, after events have already unfolded and the decisions have already been made. Consequently, a distinct subfield of research focuses on refining analytical techniques to enhance the quality of intelligence products. A prime illustration of this effort is the development of structured analytical techniques (SATs), aimed at mitigating cognitive biases such as anchoring bias, and mirror imaging (Pherson & Heuer, 2015).

CORE DISCIPLINE	SUBDISCIPLINE	EXPLANATION	EXAMPLE
GEOSPATIAL INTELLIGENCE (GEOINT)	- Imagery intelligence (IMINT) - Infrared intelligence (IRINT)	Utilizing and examining images and spatial data to characterize, evaluate, and visually represent natural attributes and activities with geographic references on or concerning the Earth.	Analysis of satellite imagery to assess the impact of natural disasters
HUMAN INTELLIGENCE (HUMINT)		Gathering and scrutinizing data from human sources, typically via direct interaction between intelligence personnel and individuals holding pertinent information. This encompasses a wide array of information acquired through various means such as interviews and espionage.	An agent builds rapport with a source within a terrorist group, gathering information about their activities and plans.
MEASUREMENT AND SIGNATURE INTELLIGENCE (MASINT)	- Radar intelligence (RADINT) - Radiation intelligence (RINT) - Nuclear intelligence (NUCINT) -Acoustical intelligence (ACINT) - Electro-optical intelligence (ELECTRO-OPTINT) - Infrared intelligence (IRINT)	Identify and examine unique physical or technical signatures associated with targets of interest. Various data collection methods, including electromagnetic, seismic, acoustic, and nuclear radiation sensing, as well as material and chemical analysis, are utilized in MASINT.	Use of radar cross-section (RCS) measurements to characterize and identify different types of aircraft, ships, or ground vehicles.
CYBER INTELLIGENCE (CYBINT)		Identify, monitor, and evaluate the digital threats posed by adversaries. These activities typically involve monitoring private networks, computers, and communication devices that are essential for day-to-day operations	Monitoring network traffic for suspicious patterns or anomalies that may indicate a potential cyber-attack.
OPEN-SOURCE INTELLIGENCE (OSINT)		Gathering, processing, analyzing, producing, categorizing, and sharing information that is sourced openly and legally accessible to the public.	Monitoring public social media posts to gather information on members of an extremist organisation
SIGNALS INTELLIGENCE (SIGINT)	- Electronic signals intelligence (ELINT) - Foreign instrumentation signals intelligence (FISINT) - Communications intelligence (COMINT)	Gathering intelligence by intercepting and analyzing various types of signals. This involves intercepting people's communications, monitoring electronic signals from different devices and systems, and analyzing signals from foreign equipment.	Intercepting and analyzing radio signals emitted by enemy radar systems to gather intelligence on their location, capabilities, and strategy.
SOCIAL MEDIA INTELLIGENCE (SOCMINT)		Systematic collection and analysis of data obtained from a multitude of social networks, channels, and communication groups.	Studying public sentiment by analyzing posts on Reddit

Table 1: Overview of Intelligence Disciplines based on Henrico and Putter's (2024) classification

After the complicated analysis stage, the next and final step is dissemination, wherein the meticulously crafted intelligence output finds its way back to the initial stakeholders (Phythian, 2013). This marks the pivotal point where policymakers, entrusted with the finished product, harness its insights to steer decision-making. It is noteworthy that subsequent policy actions may, in turn, prompt additional requisites for intelligence, thus perpetuating the cyclical nature of the intelligence process (Phythian, 2013).

2.2.1. Critiques of the Intelligence Cycle

While the intelligence cycle remains a cornerstone in intelligence organisations and educational contexts, it has faced scepticism and criticism from various authors regarding its practical application in intelligence organisations. Hulnik (2006) for example notes how the process of intelligence analysis does not happen in a clear-cut sequence as the cycle suggests. Instead, he observes that certain elements of the process occur simultaneously, particularly data collection and analysis. These two stages of the process mutually inform each other; during the analysis stage, an analyst may for example identify a knowledge gap, prompting further data collection. Moreover, Hulnik (2006) underscores that the traditional intelligence cycle fails to adequately address counterintelligence and that in practice the delineation of responsibilities between the steps is often blurred.

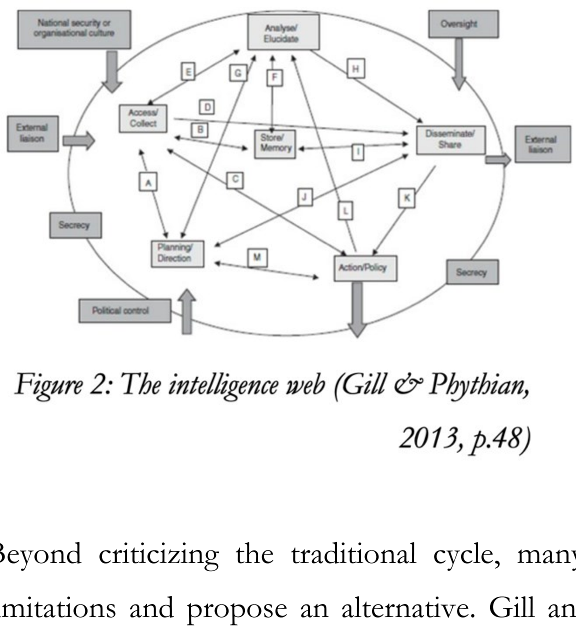


Figure 2: The intelligence web (Gill & Phythian, 2013, p.48)

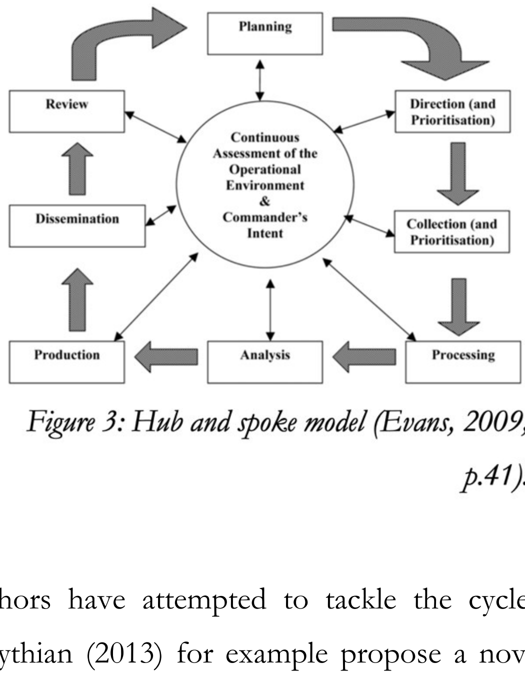


Figure 3: Hub and spoke model (Evans, 2009, p.41).

Beyond criticizing the traditional cycle, many authors have attempted to tackle the cycle's limitations and propose an alternative. Gill and Phythian (2013) for example propose a novel viewpoint to shed light on the limitations of the intelligence cycle framework, placing particular emphasis on what they name complexity challenges. The authors have outlined seven challenges, including bureaucratic dynamics, comparative assessment, interpreting intelligence via a risk-oriented lens, interactive dynamics, technological influences, covert operations, and oversight mechanisms. Their analysis rejects the traditional sequential cycle, advocating for a more nuanced and interconnected approach instead, which they term the *intelligence web* as pictured in Figure 2. Another alternative to the traditional cycle is proposed by Evans (2009), who puts forward an expansion of the Intelligence Cycle by introducing the *hub and spoke model*. As can be seen in Figure 3, this model divides planning, review, and direction into separate functions, while the steps of collection and dissemination remain unaffected. There is another cell at the centre of this model, which receives information from all stages and conducts a constant and thorough review of the entire process.

Despite the valid and well-founded criticisms levelled against the intelligence cycle, alongside some promising proposed alternatives, the traditional cycle still holds a central role within the intelligence domain (Burke, 2021). While ongoing debates persist regarding the cycle's effectiveness and the viability of alternative approaches, the intelligence cycle undeniably highlights the core elements of the intelligence process—particularly collection and analysis, as elucidated by Jensen et al. (2022). In practical application, the boundaries between steps may indeed be blurred and intertwined, thereby deviating from a strictly cyclical pattern. Nevertheless, the indispensability of data collection and analysis in producing a refined intelligence output is hard to challenge. Indeed, Brantly (2013) underscores that while emerging technologies and evolving environments, such as the cyber domain, profoundly influence intelligence practices, the cycle's fundamental objectives remain unchanged. As such, while acknowledging the limitations of the traditional model, it has continued to provide a solid foundational framework for intelligence analysis.

Still, the advent of AI might be the first emerging technology to challenge Brantly's (2013) statement. As this dissertation will demonstrate, AI's capabilities and integration are unparalleled compared to other technologies that merely assist humans with individual tasks in the intelligence cycle. Instead, AI's integration is characterised by a significant degree of agency, transforming it from a mere support tool to a colleague with a genuine say in the process. This shift is revolutionary not in transforming the issues and data that run through the cycle, but in how the cycle addresses these issues. For the first time, the core framework may therefore indeed need to be revisited to accommodate this enhanced form.

2.2.2. Automating the Intelligence Cycle

The evolution of intelligence collection disciplines vividly demonstrates the dynamic nature of the intelligence landscape, mirroring the rapid changes in the world and the ubiquitous presence of technology. Technology has seamlessly integrated into our daily routines, generating vast quantities of data commonly referred to as big data (Symon & Tarapore, 2015). At first sight, this abundance of data looks like a boon for intelligence endeavours, as more information promises enhanced insights and understanding of a particular situation. However, this relationship is much more complex than that. Beyond the surface, analysts grapple with the complexities that accompany the copious amount of data (Stenslie, 2023). The sheer volume not only increases the likelihood of encountering unreliable or misleading, information, the task of collecting, organising, and synthesising this wealth of information exceeds the capacity of human cognition (Stenslie, 2023). Simply put, while big data holds potential, it also poses significant challenges for effective analysis. The concept of cognitive capacity hereby becomes pivotal (Regens, 2019). Human brains have inherent limitations regarding the amount of information they can process. Consequently, a disparity emerges between the optimal analysis achievable and that produced by a team of human intelligence analysts (Regens, 2019). As the volume of data continues to burgeon, it becomes evident that relying solely on human efforts is no longer feasible. Turning to technical solutions therefore becomes not only preferable but seemingly inevitable.

Intelligence agencies are progressively embracing technology to automate various aspects of the intelligence cycle (Erskine & Miller, 2024; King, 2024; Logan, 2024). King (2024) for example describes how the US XVIII Airborne Corps used AI-processed data to help the Ukrainians identify Russian targets. This integration of AI into intelligence operations is further exemplified by the longstanding and evolving partnership between tech giant Palantir and the U.S. Department of Defence (DoD). Palantir for example leverages its expertise to implement automatic target recognition into the airborne systems utilised by the Army (Palantir, n.d.-b). Furthermore, Palantir's Gotham is an advanced software solution that centralises and automates intelligence analysis processes (Palantir, n.d.-a). The AI-powered tool is capable of seamlessly integrating, securing, managing, and analysing near-real-time data (Daniel, 2022). For example, in the event of enemy detection, Gotham simulates various factors such as terrain conditions, refuelling options, and air coverage to determine the optimal extraction route and the time required for the evacuation of troops (Palantir, 2020).

Palantir's partnership with the US DoD merely scratches the surface of a vast landscape of collaboration, with companies such as Airbus launching similar systems (Airbus, n.d.). Numerous cutting-edge technologies for automation are readily accessible and remarkably efficient. For instance, Hoppa et al. (2019) introduced a tool designed to automate the collection and analysis of public X data. This tool allows the analysis of thousands of tweets in a fraction of the time and cost it would take a human to perform the same task. With a plethora of similar examples (e.g., Kumar et al., 2021; Marzouk et al., 2023; Neupane & Seok, 2020), the existing body of literature predominantly discusses strategies and methodologies for automating the collection, processing, and analysis steps of the intelligence cycle. This focus comes as no surprise, as these are the steps that demand the most time, effort, and data, offering the potential for significant gains. Notably, an intriguing trend observed in the literature is the growing usage of AI to achieve such optimisation (e.g., Jiang, 2020; Li et al., 2021; Liu et al., 2021).

2.3. Artificial Intelligence

2.3.1. Machine Learning and Deep Learning

AI has emerged as a dynamic frontier in technological advancement, aiming to equip machines with human-like cognitive abilities and intelligent behaviour (Xu et al., 2021). The technical workings of AI encompass a spectrum of techniques that enable machines to quickly and accurately perform a series of tasks when provided with a prompt (Xu et al., 2021). Examples of such intelligent behaviour include speech recognition, identifying objects in pictures, or finding patterns in written text. Initially, expert systems utilised a knowledge base and inference engine for problem-solving. However, the landscape has shifted towards Machine Learning (ML), where algorithms learn and refine themselves through experience (Garnelo & Shanahan, 2019). That is, the system learns from the data, as opposed to learning from the programs that humans created.

Each ML model has an input and an output, which respectively refer to the data which is fed to the model and a particular prediction based on this data (Alpaydin, 2020). In the field of intelligence, the input could take the form of a satellite image of a particular area, while the output may involve identifying the exact location of a nuclear plant within that area. To evaluate a model's performance, a training objective is used, which serves as a guide for the learning algorithm to adjust the model parameters and optimise its performance with respect to the training data (Alpaydin, 2020; Dogo et al., 2018; Rao et al., 2023). In most cases, this involves iteratively calculating gradients of the objective function with respect to the model parameters and updating the parameter values to move toward the optimal solution (Dogo et al., 2018; Rao et al., 2023). This process of iterative optimization is commonly referred to as gradient descent. The aim of optimising the model parameters is to improve the model's ability to generalise to new, unseen data and make accurate predictions (Alpaydin, 2020).

The field of ML encompasses several approaches, of which the most common are unsupervised learning, supervised learning, and reinforcement learning (Morales & Escalante, 2022). While supervised learning uses labelled data sets and unsupervised learning detects patterns on its own, reinforcement learning uses interactive experiences to aid decision-making (Morales & Escalante, 2022). The choice of which approach to use depends on the specific application and the desired outcome. ML yields valuable results; however, for intricate decision-making, Deep Learning (DL), a subset of ML, holds even greater promise. Unlike classic ML, DL can automatically discern patterns and features from unstructured data such as text and images, without requiring a labelled dataset (Goodfellow et al., 2016). These models excel due to their ability to learn complex features hierarchically and integrate them effectively (Goodfellow et al., 2016).

2.3.2. (Convolutional) Neural Networks

DL is built upon the concept of neural networks – computational models that are designed after the workings of the human brain (Aggarwal, 2018; Goodfellow et al., 2016). Neural networks, as pictured in Figure 4, comprise node layers, which consist of input, hidden, and output layers. These layers are composed of artificial neurons, each of which has a weight and a threshold value (Aggarwal, 2018). If a neuron's output exceeds the threshold value, it activates and transmits its data to the subsequent layer of the network. Conversely, if the neuron's output falls below the threshold, no data is transmitted. The fundamental mechanics behind neural networks thus involve calculating the weighted sums of values in the input layer and then establishing a connection between the input and output layers through a sequence of typically nonlinear functions (Aggarwal, 2018; Goodfellow et al., 2016). The quantity and configuration of layers, the specifications of each individual layer, and the manner in which these layers are interconnected are what make up the neural network's architecture.

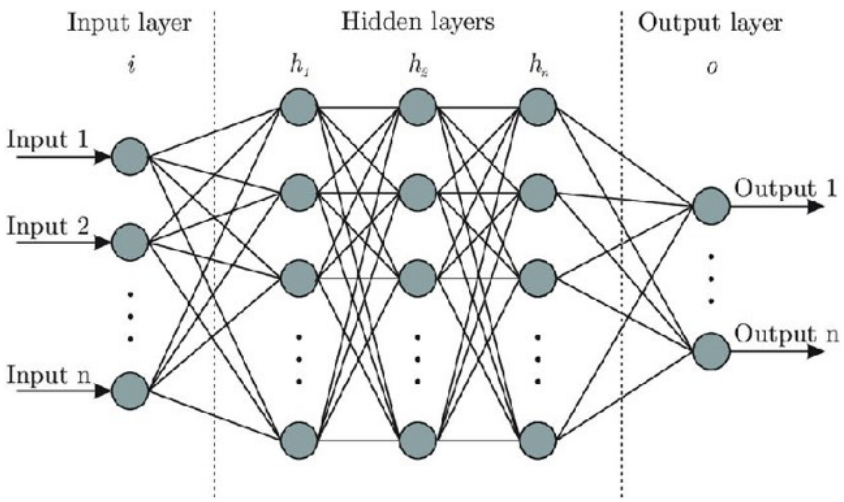


Figure 4: Neural Network Structure (Alfayoumi et al., 2021, p. 302)

Convolutional neural networks (CNNs) are a specialised type of DNN that stands out in the realm of computer vision, which deals with image classification and other types of grid data such as audio, signal, radar or sonar (Gu et al., 2018; Li et al., 2022). Although a CNN shares a similar architecture with a regular neural network, its key difference lies in the way neurons are connected as can be seen in Figure 5. In contrast to being fully connected, neurons in a CNN are not required to connect with every neuron in the previous layer (Gu et al., 2018; Li et al., 2022). Rather, only those neurons within a specific receptive field are selectively connected. Simply put, a CNN breaks down an image into smaller sections rather than analyzing the entire image at once, making it highly adept at recognizing local features (Gu et al., 2018; Li et al., 2022). For instance, when presented with an image of a car, the CNN dissects the image to identify individual elements such as wheels, headlights, windshield wipers, a license plate, a car logo, and more. It then combines these elements to conclude that these elements together are most likely a car.

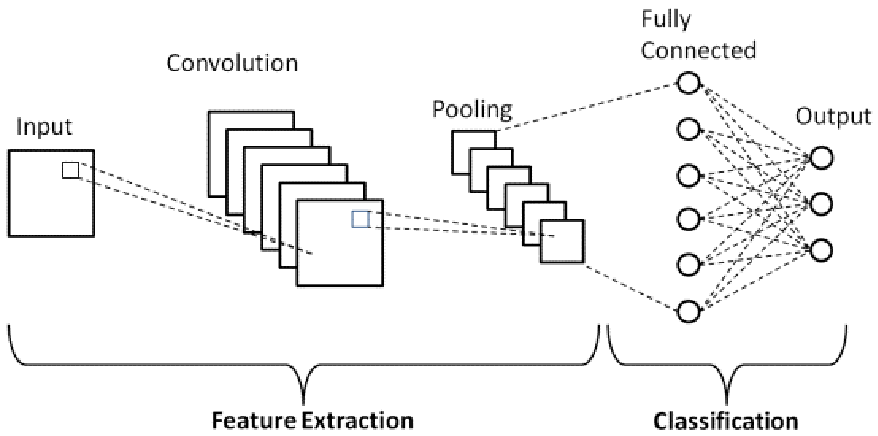


Figure 5: Convolutional Neural Network (Balaji, 2020, p. 1)

2.3.3. Generative AI, Large Language Models, and Multi-Agent Systems

DNNs have been demonstrated to be highly versatile and can be used for a wide range of applications beyond simple image recognition. Despite the existence of AI technology for some time, it particularly gained widespread popularity after the development of ChatGPT – an advanced AI-powered chatbot created by OpenAI (Hao et al., 2024; Wu et al., 2023). ChatGPT's remarkable ability to answer complex user queries has been a major milestone in the field of AI research and has put generative AI (GenAI) in the global spotlight (Wu et al., 2023). GenAI is a powerful tool that can generate hyperrealistic images, videos, voice recordings, and comprehensive analyses, making it a promising asset for decision support and automating elements of the intelligence cycle.

One particular model type used in GenAI is a large language model (LLM). LLMs represent a groundbreaking innovation in natural language processing (NLP), leveraging vast datasets and sophisticated neural network architectures to comprehend and generate human-like text (Amaratunga, 2023; Hao et al., 2024; Kaswan et al., 2023). Through iterative training stages and ongoing refinement, LLMs continue to evolve, offering immense potential in various applications, from language generation to conversational agents and beyond. At its core, LLMs involve predicting the next word based on the preceding sequence of words, essentially framing the task as a classification problem (Amaratunga, 2023). This approach enables the model to learn complex patterns and relationships within language data, resulting in coherent and contextually relevant outputs (Amaratunga, 2023; Kaswan et al., 2023). The extensive scale of parameters within LLMs facilitates their ability to learn and generate text across diverse contexts, encompassing languages, genres, and formats. While LLMs are perhaps the most popularly known form of AI, they serve as just one example within the broader landscape of AI, as depicted in Figure 6.

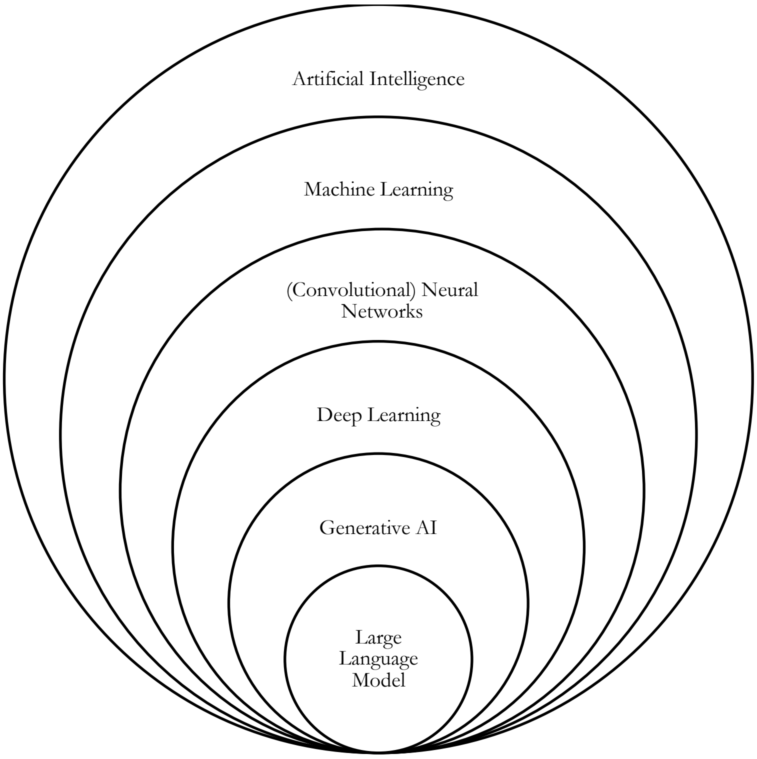


Figure 6: Relation between different AI-related concepts

While LLMs already serve as a promising tool for automating the intelligence cycle, the integration of diverse model types in combination with one another opens up even more possibilities. This approach, known as a multi-agent system (MAS) entails the orchestration of multiple decision-making agents operating within a shared environment, collaborating to achieve common or conflicting goals (Alzetta, 2020; Calvaresi et al., 2021; Calegari, 2021). Through MAS, disparate model types can synergise, leveraging their respective strengths to tackle complex analytical tasks with enhanced efficiency and accuracy (Alzetta, 2020). This collaborative framework not only amplifies the capabilities of individual models but also facilitates the emergence of sophisticated behaviours and adaptive strategies. By using MAS decision-making processes can be even more streamlined and complex datasets can be navigated with greater agility and precision (Rizk, 2018).

2.3.4. The Black-Box Problem

Despite offering great potential, AI is not without its shortcomings. The intricate neural networks, while enhancing accuracy, render AI a *black-box* technology. This means that despite highly accurate outputs, the designers of the model may not comprehend the underlying reasons for this accuracy (Ehsan et al., 2022; Eschenbach & Warren, 2021; Rai, 2019). With a vast array of nodes and connections, thoroughly analyzing each step to comprehend why, for instance, a car logo was favoured over windshield wipers in differentiating a car from a bus, becomes nearly impossible. As such, while the outcomes may be accurate, the rationale behind them may remain elusive. In the realm of intelligence analysis, this presents a significant challenge. AI might base its decisions on seemingly inconsequential factors, potentially skewing the analysis (Eschenbach & Warren, 2021). The field of explainable AI (XAI) has emerged to tackle this problem, aiming to enhance the transparency of AI decision-making processes (Ehsan et al., 2022; Eschenbach & Warren, 2021). While achieving high explainability often entails a trade-off with accuracy, much effort is being made to strike a balance (Van Der Veer et al., 2021; Vössing et al., 2022). Vössing et al. (2022) for example aim to provide human operators with insights into why the system makes certain decisions, enabling them to appropriately evaluate the system's trustworthiness while maintaining a high-quality analytical output.

2.4. Adversarial AI

The lack of explainability is not the only concern that comes with the intricate mechanics of AI. AI systems are namely prone to manipulation – known as Adversarial AI (AAI), a field of research exploring how AI structures can be manipulated for strategic advantage (e.g., Goodfellow et al., 2018; Irfan et al., 2021; Papernot et al., 2016). In contrast to the conventional approach to machine learning, which assumes a benign environment throughout the model's training and evaluation, AAI delves into the ways in which a malicious actor could manipulate the model either during the training, testing, or deployment phases (Goodfellow et al., 2018; Papernot et al., 2016). While ample literature focuses on defending ML models against adversarial attacks, for example by making the training data more exhaustive, it is crucial to recognise that adversarial examples are inherent to the technology's design as opposed to a bug that can be solved (Ilyas et al., 2019). Hence, the pattern recognition systems that contribute to the predictability of machine learning are also the primary cause of non-robust features, which, in turn, make the system susceptible to adversarial manipulations (Ilyas et al., 2019). Simply put, while defensive measures are possible and needed, the possibility of manipulation can never fully be eradicated. In contrast to flaws arising from the system's original design, vulnerabilities exploited by adversarial actors are not easily discernible through conventional testing and evaluation procedures. This unpredictability presents a considerable obstacle to formulating regulations capable of effectively safeguarding against malicious manipulations (Goodfellow et al., 2018; Ilyas et al., 2019).

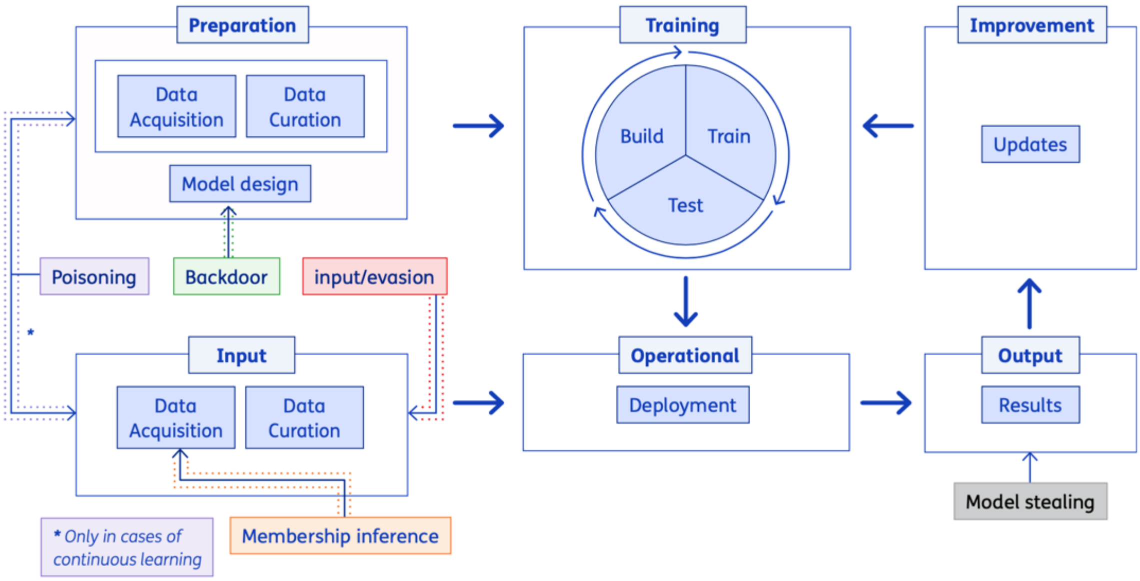


Figure 7: AAI attack categories in ML cycle (Brink et al, 2023, p. 3)

The literature discusses several attack methods for manipulating AI systems of which an overview has been made by Brink et al. (2023) as visualised in Figure 7. One example is a backdoor attack in which perpetrators introduce a piece of code into the model, lying dormant until specific conditions are met (Guo et al., 2022). Upon activation, the backdoor enables the model to generate the desired output for the attackers. Although primarily theoretical, recent research indicates the potential existence of invisible backdoors under specific conditions (Goldwasser et al., 2022). Another form of adversarial manipulation is a poisoning attack, involving the manipulation (addition, removal, or modification) of training data to heighten the probability of misclassifications by the model (Biggio & Roli, 2018). Perpetrators execute this type of attack by altering the delineations between object types in the training data, causing specific objects to be misclassified when presented to the model (Biggio & Roli, 2018). These examples highlight the wide range of attack methods that reveal the vulnerability of machine learning models at various stages of their lifecycle. During the development phase, attackers might manipulate training data or insert a backdoor, and after the model's release, attackers can exploit access through deception or even attempt to replicate the model.

Given the seemingly inevitable integration of AI in processing big data for intelligence analysis, it becomes imperative to thoroughly comprehend the risks posed by AAI on the automated steps in the intelligence cycle. Despite existing research that outlines future possibilities and complications integrating AI in the intelligence process (Sadiku & Musa, 2021; Blanchard & Taddeo, 2023), a significant research gap exists regarding the specific risks of AAI, which could have profound implications for the quality of intelligence products and, consequently, national security. Indeed, if the limitations of AI are already identified as a national security concern by Erskine and Miller (2024) as well as Logan (2024) without considering manipulation – introducing AAI into the equation would only work to further exacerbate these concerns. This dissertation aims to further explore the implications of this added factor, providing essential insights into what to consider when implementing AI to automate the intelligence cycle.

Chapter 3

Methodology

To ensure the safe implementation of AI, it is crucial for intelligence agencies to explore the threat posed by AAI to the effectiveness of the intelligence cycle. This understanding will allow them to adjust their existing frameworks and strategies accordingly. While of great importance, it is not without reason that this area remains underexplored as this topic comes with several inherent research constraints. Firstly, AI automation is still in its early stages. This is not only because the computational power required for AI to match or transcend human proficiency was only reached fairly recently, but also because in the realm of national security, even minor errors can lead to significant consequences. As a result, secure deployment requires meticulous and time-intensive procedures (Intel AI, 2019). Additionally, while the AI research field is extensive and rapidly developing, studies on AAI and the impacts thereof are far from reaching the same level. In scholarly discussions, attack methodologies are frequently theoretical or tested in the most realistic setting possible. As such, while AAI taxonomies and surveys highlight the increasing realism of adversarial attacks (e.g., Vassilev et al., 2024; Hossain et al., 2024), real-world examples and use cases remain scarce. This scarcity may stem from a delay between academic discoveries and practical implementation, or from victims either failing to detect infections or refraining from reporting them. Lastly, the practical application of the intelligence cycle occurs in highly secretive and confidential environments. As such, obtaining primary real-life data on how the cycle is automated, how these modern technologies compare to traditional means, and how adversarial attacks influence the ultimate intelligence product, is exceptionally challenging.

Even though it is crucial to assess the risks of AAI to automated intelligence analysis, as it concerns matters of national security, there is a gap in the literature that examines these threats due to the aforementioned constraints. To address these challenges and bridge the gap, this dissertation aims to utilise state-of-the-art literature on AI and AAI to analyse their impact on the traditional intelligence cycle and propose adjustments to optimise the cycle accordingly.

3.1. Research Design: Case-Study

To address the dissertation's complex research question and provide actionable insights into updating the intelligence cycle in the face of AAI threats, a case-study approach has been selected. A case study involves an in-depth examination of a particular case or a few cases, using observational data to provide insights into a larger population of cases (Gerring, 2017, p. 28). This method is characterised by its high level of focus, as it involves analysing and presenting the selected case(s), using the details of the case to support the overall argument (George & Bennet, 2005; Gerring, 2017). Furthermore, George and Bennet (2005) note how a major advantage of case studies is they help achieve high levels of conceptual validity, as the concepts are discussed and analysed in the appropriate context.

An *instrumental* case study is a specific type of case study approach that explores several specific instances that provide insight into a particular area or phenomenon more than the individual cases themselves (MacQuarrie, 2010; Stake, 2005). In essence, the aim is thus to study a sample, consisting of several cases, that together, to some extent, represent a larger phenomenon. This approach can help make existing knowledge more workable and explicit (MacQuarrie, 2010). An instrumental case study is therefore a fitting method to address the dissertation's research question, as the little data that *is* available, can be utilised in such a way that it bridges an important research gap.

3.1.1. Case Selection Criteria

To explain the broader phenomenon of how AAI influences the intelligence cycle and how it should be revisited accordingly, the cases selected for the sample should explain a core part of the issue. These cases should demonstrate how individual threats impact the overall usability and effectiveness of the intelligence cycle. By doing so, they will provide critical insights into the specific ways AAI shapes and potentially disrupts the intelligence cycle, highlighting areas that require re-evaluation and adaptation. The method best suited to achieve this aim is what Seawright and Gerring (2008) term the *diverse case method*. The goal of this sampling method is to obtain the highest possible variation in important dimensions (Seawright & Gerring, 2008). Compared to other case study sampling methods, the diverse case method is most apt to be representative of the broader phenomenon under study. While it is essential to acknowledge that these cases may not perfectly reflect the distribution of this variation in the population, they go a long way in achieving a comprehensive understanding of the population (Seawright & Gerring, 2008).

As the aim is to comprehend the impact of AAI on the entire intelligence cycle's usability, it is necessary to include representative cases that cover different stages of the cycle as well as various forms of AAI. Dedicating each case to a particular stage of the cycle should provide a detailed yet holistic analysis. Moreover, to gain a more comprehensive understanding of the broad scope of AAI, each case should also consider a different type of AI application, as different applications and models come with different attack vectors.

To ensure a representative sample, five cases corresponding to the five steps in the cycle were initially considered. However, increasing the number of cases simultaneously leads to a decreased level of detail and attention each case receives (Gerring, 2017). Since each step must incorporate AI automation, address complex attack types, and connect it to the intelligence cycle, focusing on all five steps would diminish the much-needed specificity. Therefore, to ensure population representativeness and achieve the necessary depth, the focus will be on the steps in the cycle most likely to be automated or most influential to the final output: collection, processing, and analysis.

Jensen et al. (2022) posit that the essence of the intelligence cycle resides in its pivotal stages: collection and analysis. These stages lay the groundwork for the entire process, with the processing phase serving as the vital link between them, assuming a significant role by default. Not only do these stages shape the essence of the final intelligence output, but they also demand the most labour from analysts, who are expected to surge through vast data repositories. Given AI's ability to swiftly process massive data volumes and discern intricate patterns within them, there exists great potential for streamlining these critical stages. Consequently, automating precisely these processes could yield substantial benefits. This also implies that there will likely be a greater reliance on AI in these steps. Therefore, studying these specific cases provides the most profound insight into how the traditional cycle needs to be adjusted.

To assure optimal diversity and representativeness, every case will explore a distinct approach to AI. This diversity must be reflected either in the type of AI model used or the context of application, allowing enough flexibility for the case to examine the most suitable and practical application for its respective step in the cycle.

3.1.2. Selected Cases

Based on the selection criteria and literature review, three cases have been selected as a sample to represent the broader population, as summarised in Figure 8.

Case 1 investigates the potential ramifications of AAI in the field of Geospatial Intelligence (GEOINT). The use of CNNs in computer vision has proven to be an efficient tool for identifying the presence, absence, and classification of objects of interest. This has resulted in the development of autonomous drones that can collect data without requiring human intervention (Jouhari et al., 2022). These unmanned aerial vehicles (UAVs) are more cost-effective and safer compared to deploying human operators in hostile environments. However, AAI can influence the data collected by these autonomous drones, compromising their reliability and usefulness. Notably, evasion attacks, which are extensively documented in the realm of computer vision, emerge as a salient threat vector warranting attention in this context (Qian et al., 2020).

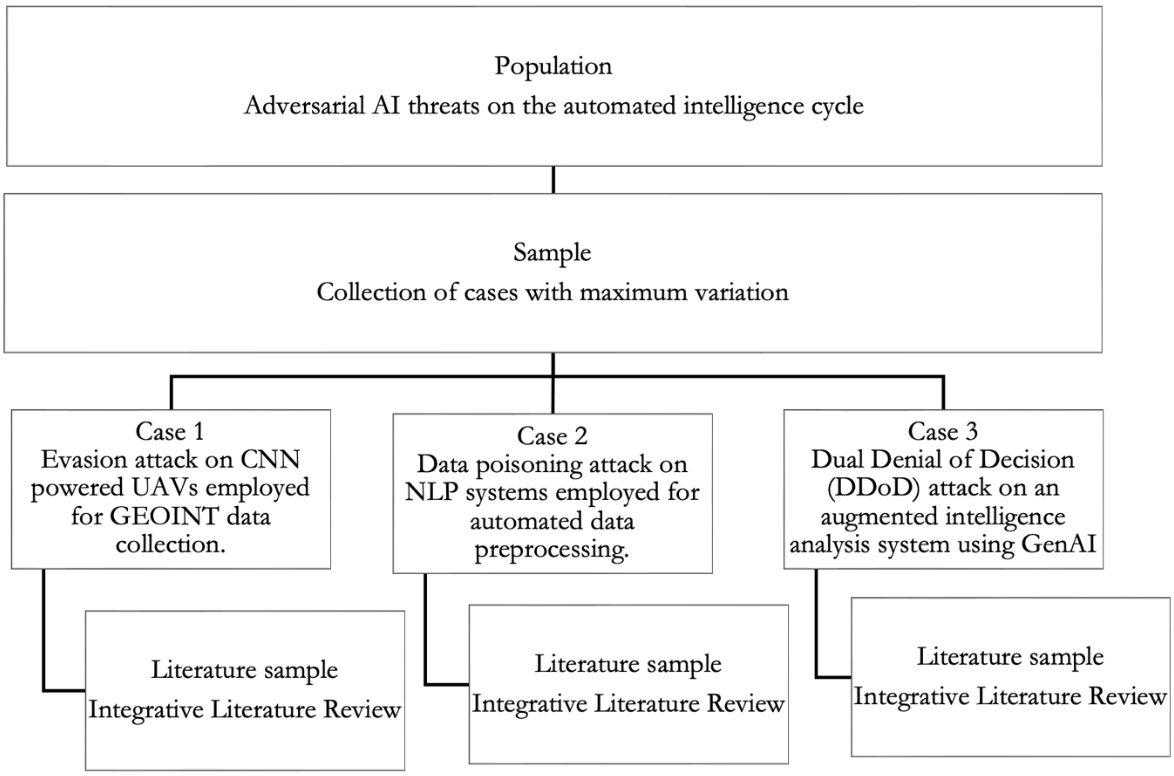


Figure 8: Overview of Research Design

Case 2 delves into the automation of the processing stage within the intelligence cycle, focusing on the integration of Natural Language Processing (NLP) to streamline data preprocessing. Raw data requires preprocessing before it can be effectively comprehended and utilised by humans, involving tasks such as translation, error detection, consistency checks, and misinformation identification (Christa et al., 2013). Given the varying reliability of data sources, NLP enables AI to swiftly navigate extensive textual datasets, identifying errors, outliers, and categorising data to enhance usability (Billal et al., 2016). Consequently, it holds significant promise in improving the accuracy and efficiency of data cleaning and transformation compared to manual efforts. However, the threat of adversarial manipulation poses substantial challenges, particularly for analysts reliant on precise data for their analyses. Poisoning attacks, which manipulate the training data rather than the input data as seen in Case 1, emerge as particularly intriguing in the context of data preprocessing (Marulli et al., 2021). A successful poisoning attack can compromise the integrity of the data processing stage, directly opposing its intended function. This highlights the critical need for further investigation.

Case 3 scrutinises the convergence of GenAI and intelligence analysis. GenAI systems can empower intelligence analysts by using numerous datapoints to generate potential Courses of Action (COAs), providing valuable insights for proactive strategic and tactical analyses to support subsequent decision-making (Erskine & Miller, 2024; Schwartz et al., 2020). This integration typically adopts a hybrid approach, where a human analyst remains integral to steer and validate AI-generated outputs, ensuring maximal accuracy and relevance (Wu et al., 2022). As such, traditional attacks targeting only the AI-enabled entity, which dominate current AAI literature, are not successful by default. Hence, beyond merely fooling the system, a successful attack now also needs to fool the human. Moreover, adding a human-in-the-loop even adds another non-technical attack vector: the human analyst. The Dual Denial of Decision (DDoD) attack is unique in this context, as it exploits this dynamic by deceiving both the AI system and manipulating the cognition of the human analyst (Tag et al., 2023). This type of attack has significant implications for the analysis process and therefore provides a realistic attack method to investigate.

3.2. Data Collection

Studying intelligence-related matters presents profound challenges due to national security concerns, often resulting in the classification and limited accessibility of relevant data (Kerr, 2023). Furthermore, the early stages of AI implementation in intelligence analysis mean that primary data on its effectiveness and application is either unavailable or restricted (Intel AI, 2019). To overcome these constraints, this dissertation utilises existing literature to draw novel conclusions. Fortunately, there is a wealth of up-to-date and detailed information on AI and adversarial attacks available in academic literature, conference papers, and reports from major tech companies. However, while this extensive body of literature provides a solid foundation for identifying and understanding the threats that need to be addressed before fully implementing AI technology—

such as Logan's (2024) exploration of using LLMs for intelligence analysis—it is rarely connected to its impact on the entire intelligence cycle. This gap highlights the need for focused research on how AAI specifically affects these critical areas, and how the intelligence cycle must be reimagined to accommodate the fundamental changes brought about by AI-integration.

To ensure the quality and relevance of the analysis, several selection criteria have been established. Only peer-reviewed Q1-ranked journal articles, conference papers, and reports published within the last five years are considered. This focus ensures that the studies have undergone rigorous evaluation by experts in the field and are among the top-tier publications, providing reliable and significant contributions to the subject matter. Moreover, limiting the scope to publications within the last five years helps ensure that the analysis is based on the most up-to-date information available, reflecting the current state of knowledge and state-of-the-art practice in the field. Furthermore, articles lacking empirical data or rigorous methodology are excluded. Empirical data provides concrete evidence to support research claims, while rigorous methodology ensures that studies are conducted in a systematic and controlled manner. Overall, these selection criteria collectively contribute to the credibility and validity of the analysis by prioritising high-quality, relevant, and recent research findings.

The data collection process for this study involves a systematic search strategy targeting academic literature, conference papers, and reports from major tech companies. The consulted database was the Scopus database due to its extensive and interdisciplinary respiratory, containing entries from high-quality databases such as ACM Digital Library and IEEE Xplore. ArXiv was also considered due to its relevant body of computer science literature, however as not all articles are peer-reviewed, this database was purposefully left out. To yield as many relevant results as possible for each case, several relevant keywords were identified. These keywords and the search queries used for each database and case can be found in Appendix A.

All the search results were downloaded and then imported into ASReview, an AI-powered tool that significantly enhances the systematic literature review process. Unlike traditional manual methods, ASReview allows researchers to efficiently filter through large databases by rapidly assessing titles and abstracts against pre-defined inclusion criteria. This approach allows for the consideration of a greater number of articles and accelerates the identification of the most pertinent literature, ensuring the selection of high-quality, relevant sources crucial for effectively mapping out the threats associated with AI and adversarial attacks in automation.

The review process involves the researcher marking each source as relevant or not. As these assessments are made, ASReview's learning algorithm adapts and begins to prioritise and present the most relevant pieces of literature. For this dissertation, the evaluation process continued until ten consecutive articles were deemed irrelevant, indicating that saturation had been achieved. Following this, a thorough evaluation was conducted on the full-text articles to assess their methodological rigour, theoretical relevance, and empirical contribution.

3.3. Data Analysis

Data analysis is crucial in uncovering valuable insights and patterns from collected literature. During this stage, the aim is to combine information from selected articles to comprehensively understand the risks associated with AI and AAI to the functioning of the intelligence cycle. Specifically, the analysis is to be conducted following the principles of an *Integrative Literature Review* (ILR) – "a distinctive form of research that uses existing literature to create new knowledge" (Torraco, 2016, p. 404)

ILRs are particularly effective in synthesizing diverse methodologies and offering comprehensive perspectives on a given topic (Lubbe et al., 2020; Torraco, 2016). This makes the ILR approach especially relevant for addressing this dissertation's research question. Rather than analysing and combining primary research results, the ILR enables the acquisition of fresh perspectives and consolidated knowledge on the subject, utilising both quantitative and qualitative data (Lubbe et al., 2020). This is essential for bridging the research gap discussed in this thesis, which spans various domains including the technical nature of AI applications, adversarial attack methods, and intelligence studies. Indeed, since obtaining primary data is not feasible yet, conducting an ILR leverages existing knowledge to assess future threats that need to be understood today (Torraco, 2016). As such, the ILR method enables the contextualisation of findings from various studies within the broader framework of intelligence analysis and AAI. This is essential for understanding the practical implications of theoretical models and attack methodologies discussed in the literature, thereby providing actionable insights for intelligence agencies.

Before analysing the collected documents, a data synthesis phase was conducted. This involved amalgamating outcomes, evaluating the credibility of supporting evidence, scrutinizing the uniformity, and pinpointing any contradictory findings. Specifically, the analysis begins by categorising the literature according to fundamental themes and topics elucidated in the selected body of work. The ILR method facilitates the identification of common themes, patterns, and trends across different studies, which is essential for understanding the broader impacts of AAI and developing a coherent narrative about its influence on the intelligence cycle. These themes encompass diverse aspects but, most importantly, they help delineate different subtypes of adversarial attacks, their potential impact, and their feasibility.

Subsequently, the ILR encourages a deeper exploration into the substance of each article. This involves discerning the primary arguments and, more importantly, synthesizing them.

> Synthesizing the literature means that the review weaves the streams of research together to focus on core issues rather than merely reporting previous literature. Synthesis is not a data dump. It is a creative process that generates a new model, conceptual framework, or other unique conception informed by the author's intimate knowledge of the topic. (Torraco, 2005, p. 362)

The ILR approach therefore supports synthesising findings from high-quality studies assigning meaning in the appropriate context. This enables the dissertation to offer well-grounded suggestions for mitigating the risks posed by AAI and improving the overall effectiveness of the intelligence cycle. This process is repeated for each case to ensure consistency in results. The analysis focuses on identifying the most common themes in each case and the extent to which an attack can influence the respective step of the intelligence cycle. The findings of each case are discussed in the results section. These individual case findings are then generalised to the broader population in the discussion section, enabling the development of a revised intelligence cycle tailored for human-AI collaboration based on the insights gained from the cases.

3.4. Ethical Considerations

Regarding the data used in the dissertation, there are minimal ethical risks involved. This is because there are no participants involved, and no confidential information is handled. Still, it is important to emphasise the need for cautious presentation of findings, due to potential national security implications. The aim of the dissertation is hence to contribute responsibly to discussions about the integration of AI in the field of intelligence.

3.5. Limitations

This research design considers the availability of literature while covering diversity in AI applications and associated AAI attack types. This maximum variation increases generalisability, yet as AI is anticipated to wield influence in other steps of the intelligence cycle as well, these cases are not exhaustive. Additionally, as the AI and AAI research field is advancing rapidly, the representativeness of the case selection and the selected literature might decrease rapidly as new attack and defence methods surface. Furthermore, much like the existing body of information out there, this dissertation remains theoretical, meaning that the results might change once real-life empirical evidence is collected. It is therefore important that the dissertation's findings are constantly re-evaluated against the latest AI developments and implementation phases. Despite these limitations, the case selection realistically and impactfully represents the issue, providing a strong foundation for assessing the risks of AAI on the intelligence cycle and translating these insights into an enhanced version thereof.

Chapter 4

Results

4.1. Case 1: Evasion Attacks on CNN-Powered UAVs Employed for Data Collection

4.1.1. Automating Data Collection

The foundation of any robust analysis lies in the reliability of the underlying data. In the context of the intelligence cycle, the data collection phase is crucial, as it directly impacts the accuracy and efficacy of subsequent stages, including preprocessing and analysis (Jensen et al., 2022). Specifically, if the collected data is incomplete or contains false information, the final intelligence product will be of subpar quality. However, with the rise of big data, the issue of data scarcity has seemingly diminished, as there is now an abundance of information available on nearly everything and everyone (Symon & Tarapore, 2015). However, this shift introduces a new challenge: efficiently filtering through the vast amounts of data to find the needle in the haystack. Given limitations in time, resources, and human cognitive capacity, it is essential to employ efficient methods for collecting the most relevant and representative data (Buchanan & Kock, 2001).

Among the various forms data can take—from personal diaries to CT scans—geospatial intelligence (GEOINT) is particularly crucial for a wide range of intelligence-related issues (National Geospatial-Intelligence Agency, 2018). GEOINT is namely important for establishing situational awareness, monitoring environmental changes, and detecting objects and individuals (Utomo et al., 2021). For instance, satellite imagery can be vital for pinpointing the locations of nuclear facilities or tracking military troop movements. Yet, extracting pertinent information from this vast pool of data presents a significant challenge. Indeed, while collecting all available satellite imagery is relatively straightforward, having live footage of every inch of the globe does not simplify a robust analysis thereof. In a similar vein, Richelson (2007) noted that while technically obtained data expands the pool of information, it does not ensure that all the collected data is relevant or desired. As such, in order to maximise the efficacy of data collection, data collectors aim to gather only the most pertinent and potentially valuable information (Richelson, 2007). The integration of AI technologies can effectively mitigate this issue of information overload and optimise the data collection process (Darnell & Gopalkrishnan, 2023). That is, by harnessing AI to ensure that only the most relevant data is collected, one prevents an abundance of information from leading to decision paralysis.

Over the past decade, researchers have demonstrated that Unmanned Aerial Vehicles (UAVs) can achieve superior results in aerial data collection compared to other methods (Huang et al., 2021). UAVs can be controlled remotely, are cost-efficient, and face minimal physical obstacles during flight. The integration of ML into UAVs has even further expanded their potential for data collection (Huang et al., 2021). With the help of CNNs, these UAVs can autonomously perform sophisticated tasks such as object detection and tracking (Aposporis, 2020). Compared to traditional surveillance cameras and Low Earth Orbit satellites, AI-enhanced UAVs provide more flexible and detailed imagery (Emilien et al., 2021). Hence, they can autonomously survey areas, using AI to detect, track, and classify objects of interest from various distances.

This capability holds great potential for optimising the data collection phase within the intelligence cycle by systematically gathering the most relevant data and thereby reducing information overload, ensuring that only the most critical data is processed and analysed in the subsequent stages of the intelligence cycle. The potential of UAV technology is further amplified by advancements in swarming technology, where multiple drones communicate and collaborate to optimise their collective efforts toward a specific objective, such as gathering the most pertinent data (Zhou et al., 2020). Yet, despite offering great prospects, scientific literature simultaneously shows that such AI-powered data collection is not foolproof and can in fact be successfully manipulated, which poses significant challenges for automating the intelligence cycle.

4.1.2. Evasion Attacks in the Physical World

In the realm of image classification and object detection tasks, predominantly powered by CNNs, a notable type of attack extensively explored in the literature is the evasion attack (e.g., Ayub et al., 2020; Doss & Gunasekaran, 2023). This type of attack involves introducing specific, malicious perturbations to the input data, causing the model to misclassify the input. Various types of evasion attacks exist, depending on the goal, data type, and model used. These attacks directly threaten the integrity of the desired data and can deceive the model trained to efficiently determine which data to collect.

In the context of object detection or classification, an evasion attack is executed by altering a particular image to produce erroneous output from the CNN. Although this goal may appear difficult, numerous studies have demonstrated that CNNs can be deceived by modifying as little as a single pixel (e.g., Nam et al., 2024; Nguyen-Son et al., 2021; Yuviler & Drachsler-Cohen, 2023). These so-called one-pixel attacks, despite their deceptively minimalist nature, can lead to critical misclassifications. For example, as shown in Figure 9, altering a single pixel in an image of a horse led the CIFAR-10 model to misclassify it as a frog with 99.9% certainty.

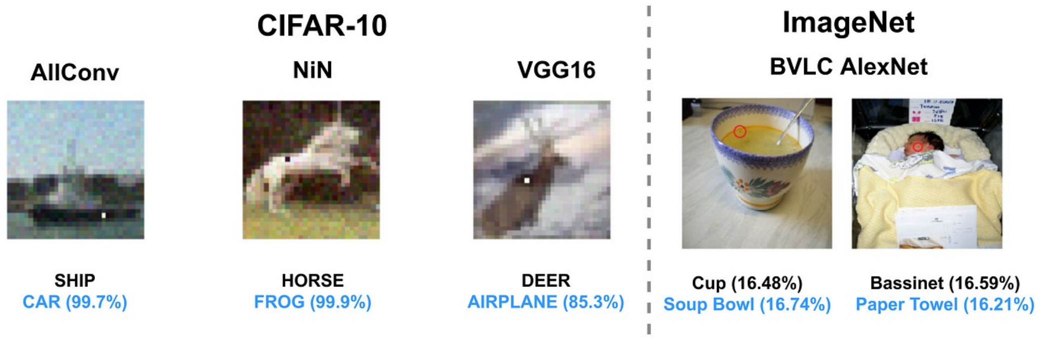


Figure 9: Example of misclassifications after a one-pixel attack. Classes displayed in black, and blue indicate the labels predicted before and after the attack respectively (Gandham, 2020, p.1).

While impactful, these perturbations have to be applied in the digital domain after an image has already been captured, ultimately leaving the physical world unchanged. This renders pixel perturbations inapplicable for real-time image or video capturing, as there is no opportunity for manipulation in the digital domain between the CNN and real-world observations. Consequently, the deployment of UAVs for automated data collection would remain largely unaffected by such attacks, preserving the integrity of the collected real-time footage.

Nonetheless, extensive literature demonstrates that the principles behind these digital pixel attacks can be adapted to work in the physical domain as well (e.g., Hu et al., 2021; Tan et al., 2021; Wei et al., 2023). That is, similar to digital one-pixel edits or hue alterations, one can introduce adversarial patches, modify textures, add 3D objects, and manipulate lighting and shadows to deceive object detection systems in real time. Given that UAVs often operate with live footage, the most pertinent adversarial perturbations discussed in academic literature are thus those occurring in the physical domain. As noted by Ojaswee et al., "Physical adversarial attacks are easier to deploy in the real world and yield higher attack success than digital perturbations" (2023, p. 4427). As a result, such physical attacks have the potential to disrupt the data collection stage, potentially causing automation to exacerbate the very problems it was designed to solve.

When examining the existing literature on evasion attacks in the physical domain, a general distinction can be made between touchable and non-touchable perturbations. Within this broad classification, numerous subcategories and specific attack methods exist, each with unique strengths and weaknesses, making them suitable for different purposes. Three types of attacks seem particularly applicable for manipulating autonomous UAVs' object detection and classification systems: adversarial patches, adversarial textures, and optical-based attacks.

4.1.2.1. Adversarial Patches

Adversarial patches, a notable type of physical adversarial perturbation, have gained significant attention in academic research, especially concerning their implications for autonomous vehicles (Eykholt et al., 2018). Yet, this concept is applicable to many objects that an adversary might seek to misclassify. For instance, an adversary might aim to keep a missile launch system undetected as it gets ready to fire at its target. In this case, traditional camouflage techniques, such as brown-green prints designed to blend with natural, bushy surroundings, will likely prove inadequate. In contrast, applying an adversarial patch to the roof of the vehicle might be more effective in hiding it from autonomous object detection. Figure 10 illustrates a compelling example from Den Hollander et al. (2020), demonstrating how adversarial patches applied to a set of fighter jets rendered them completely undetected by object detection systems. Specifically, the application of these patches reduced the model's confidence levels to between 0.01 and 0.04. Although these jets remain clearly visible to the human observer, to the object detection system these jets are nowhere to be found. Such manipulations can severely impact the data collection stage by ensuring that critical threats remain undetected.

Figure 10: Adversarial camouflage of fighter jets with adversarial patch (Den Hollander et al., 2020, p. 6)

Beyond demonstrating the general effectiveness of these attacks, research has increasingly focused on developing algorithms and other means to create adversarial patches that are more aggressive, smaller, transferable, and overall imperceptible (e.g., Lian et al., 2022; Wei et al., 2023; Yang et al., 2020). This involves determining the optimal pattern, colour, size, and shape for maximum effectiveness. For example, Wei et al. (2023) created an adversarial patch that fooled object detectors in both natural light and infrared conditions. Their approach specifically involved calculating the most effective shape and designing the patch accordingly. Figure 11 shows that the adversarial patch, consisting of two blot-shaped figures placed on a t-shirt, successfully camouflaged the person from detection. The object detector, previously successful in all instances, failed to detect the individual with the adversarial patches. Notably, the patches maintained their effectiveness across various angles, surroundings, and distances.

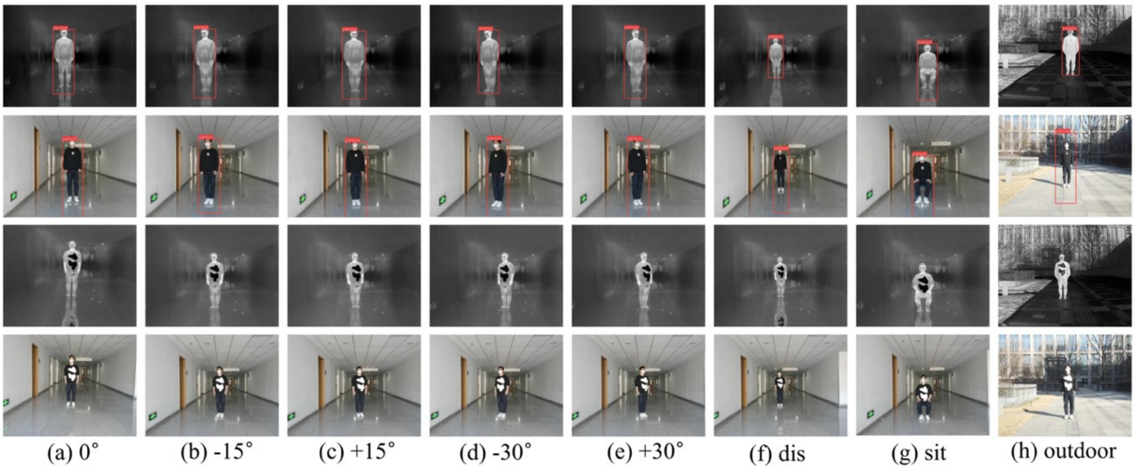


Figure 11: Adversarial patch on a t-shirt camouflaging a person (Wei et al., 2023, p. 4452)

Optimising such patches requires a thorough understanding of the object detector's mechanisms, often achieved through Gradient-Weighted Class Activation Maps (Grad-CAMs) which highlight the areas of an image attracting the model's focus (Yin et al., 2023). Hence, by identifying the areas where the model concentrates most, you can effectively determine the optimal places to modify or draw attention away from. While this suggests that attacks might only be successful in a white-box scenario—where the attacker has full knowledge of the targeted classifier—practical evidence indicates that these patches are highly transferable. This means that if an adversarial patch works on one model, it is likely to be effective on others as well. The ongoing research aims to refine these patches, making them more effective and adaptable to different detection models (e.g., Chen et al., 2024; Lian et al., 2022).

In addition to optimising the effectiveness of patches in tricking object detectors, some researchers are focusing on making these patches appear more natural to be less noticeable to the human eye (Hu et al., 2021; Hu et al., 2023; Tan et al., 2021). Indeed, while these adversarial attacks can be highly effective in fully automated data collection scenarios, they become less so when humans are involved, as people can easily spot these unusual patches. This human awareness can quickly lead to the development of countermeasures or alternative data collection methods. Consequently, attacks that blend in seamlessly and are not easily identified as malicious are considerably more threatening, posing a significant challenge to maintaining robust and reliable data collection.

Image

Setting	Lab	Living room	Hallway	Balcony	Grass
	Detection Recall				
w/o Adversarial Shirt	100%	100%	100%	100%	100%
w/ Adversarial Shirt	23.80%	24.49%	38.46%	44.33%	43.14%

Figure 12: Detection recall of adversarial example compared to regular circumstances (Hu et al., 2021, p. 7834)

Hu et al. (2021) developed the Naturalistic-Adversarial-Patch (NAP), which uses a pre-trained generative adversarial network to design patches that blend in with their surroundings. As shown in Figure 12, the detection recall[1] for a person wearing an adversarial shirt is significantly lower than for a person wearing a regular t-shirt. What makes these patches particularly intriguing is that they not only successfully deceive the classifier but may also appear inconspicuous to an unsuspecting human observer, who might perceive the adversarial example as merely a regular t-shirt with an image of a dog.

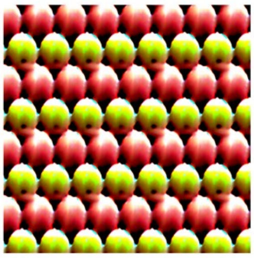
4.1.2.2. Adversarial Textures

Adversarial textures represent another frequently examined method of adversarial camouflage in the academic literature (e.g., Hu et al., 2022; Sun et al., 2023; Wang et al., 2021). A noted limitation of adversarial patches is their potential lack of visibility from different perspectives, long distances, or partial obstruction, which can diminish their efficacy. To address this issue, researchers have proposed wrapping the entire object or part thereof in an adversarial texture or pattern. This wrapping ensures that the adversarial pattern is visible from all viewpoints, thereby increasing the likelihood of deceiving the detector. As a result, it has more significant implications for

[1] *Recall* is a metric used to assess the rate at which an ML model correctly detects true positives from the total number of actual positive cases in the dataset.

autonomous UAV data collection, enhancing the effectiveness of adversarial perturbations and posing greater challenges to reliable data collection and the subsequent steps in the intelligence cycle.

Hu et al., (2022) provide a specific implementation of this attack. Instead of placing a single image on the front of a t-shirt, as was the case in the NAP, the authors use an Adversarial Texture (AdvTexture) that covers all sides of the t-shirt to avoid person detection from all angles. To ensure maximum effectiveness for this multi-angle attack, the authors have developed a "Toroidal-Cropping-based Expandable Generative Attack (TC-EGA)" to design an arbitrary yet repetitive pattern, as shown in Figure 13a. Figure 13b illustrates the effectiveness of the TC-EGA attack compared to other adversarial prints on the YOLOv2 network. The TC-EGA attack demonstrates a generally low recall, even at low confidence levels. Notably, when the model's confidence reaches 0.8, the recall drops to nearly zero. This suggests that the YOLOv2 network, when subjected to the TC-EGA attack, becomes highly confident in its classifications, despite actually encountering false negatives.



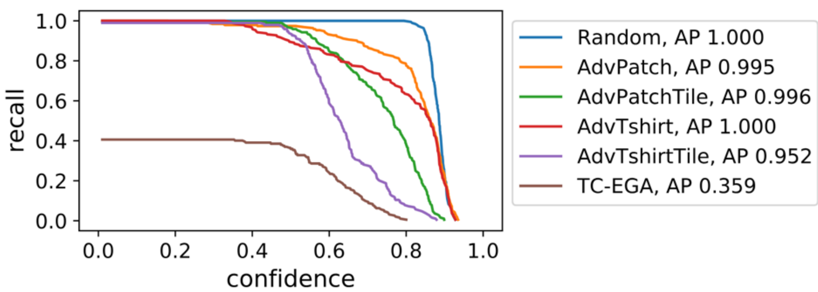


Figure 13a: TC-EGA generated pattern (Hu et al., 2022, p. 13312)

Figure 13b: "The recall vs. confidence curves and APs on the the physical adversarial test set" (Hu et al., 2022, p. 13314)

Wang et al. (2021) propose a comparable approach termed the Full-Coverage Camouflage Attack (FCA). By optimising the camouflage pattern through a loss function and applying it to the entire 3D vehicle, their technique demonstrates superior performance compared to other similar methods. Notably, the FCA is effective from multiple angles, over extended distances, and even when the vehicle is partially obscured, as shown in Figure 14. Moreover, this attack method is not only applicable to a diverse range of objects and environments but also shows consistent success across different types of model detectors.

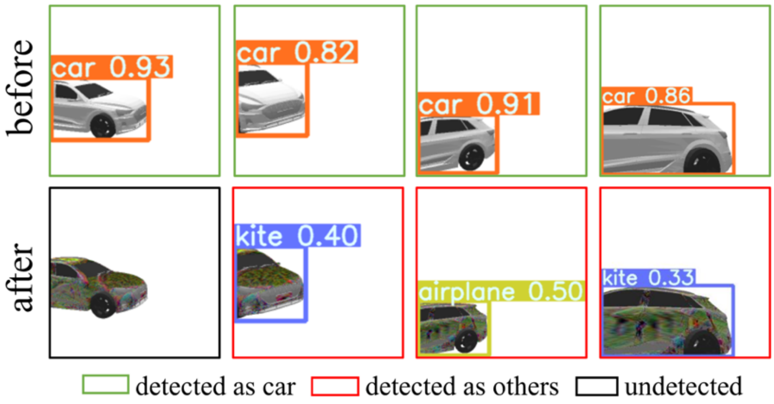


Figure 14: Success rates of FCA with partial occlusion (Wang et al., 2021, p. 2420)

While the FCA is already significantly more transferable and robust compared to a singular patch, it is still evident to the human eye that the car has been tampered with. To address this issue, Sun et al. (2023) developed a "differential evolution-based dual adversarial camouflage (DE_DAC) method," which is aimed at fooling both DNNs and the human eye. Whereas an object detector can be deceived with minor perturbations, deceiving human observers requires the object to blend seamlessly with the background. To achieve this dual deception, the authors designed a two-stage method. In the first stage, global textures are optimised to blend the object with its background, effectively fooling human observers. In the second stage, local textures derived from the global ones are refined using three loss functions to optimally deceive object detectors. The global textures are adaptive to their respective environments, while the local textures maintain consistent.

Figure 15: DE_DAC applied in various environments (Sun et al., 2023, p. 269)

Additionally, the authors employ a differential evolution algorithm to identify optimal attack regions on the object, further enhancing the camouflage's effectiveness. While the attack success rate (ASR) on object detectors tends to decrease as the naturalness of the camouflage increases, the authors achieved a suitable balance between the two. This balance resulted in the DE_DAC attack attaining an ASR between 75% and 90%. As illustrated in Figure 15, objects enveloped in DE_DAC are exceptionally well camouflaged across diverse environments, rendering them challenging to detect by both DNNs and the human eye. This dual effectiveness is particularly concerning for UAV-based data collection, as it complicates the identification and mitigation of adversarial threats. Since the camouflage is designed to deceive both DNNs and human observers, even human involvement in the data collection process might not reliably detect these manipulations. Consequently, the accuracy of threat assessments, situational awareness, and strategic decision-making processes are undermined (Richelson, 2007). This could result in flawed intelligence, and thus misinformed operations or potential failures in mission-critical activities based on inaccurate or deceptive UAV-collected information.

4.1.2.3. Optical-Based Attack

While generating and printing adversarial camouflages is highly effective, it is also extremely time-consuming and costly. For adversarial patterns, one must design an appropriate pattern and apply it to the object, which risks becoming obsolete with the deployment of new state-of-the-art object detectors and defences. Consequently, several researchers have investigated using lighting and projectors to transmit adversarial signals without altering the object itself (e.g., Chen et al., 2021; Gnanasambandam et al., 2021; Wen et al., 2023).

Figure 16: OPAD attack on a traffic sign (Gnanasambandam et al., 2021, p. 93)

Gnanasambandam et al. (2021) introduced the Optical Adversarial Attack (OPAD), which aims to mislead object classifiers without changing the physical surface of the object. Instead, a projector is used to project structured illumination on an object to give it a different appearance. While this attack has significant limitations—such as dependence on the structure and colour of the object, as well as environmental lighting conditions—it presents an interesting avenue for exploration. The authors' experiments demonstrate the attack's effectiveness for both targeted and untargeted attacks (Gnanasambandam et al., 2021). Figure 16 illustrates how an object classifier mistakenly identified a stop sign as a Speed 30 sign, with a lighting perturbation hardly visible to the human eye. This opens up the possibility of flexibly perturbing objects and adjusting these perturbations to bypass new defence methods employed in the data collection step. Additionally, when using projections an adversary can strategically decide when to activate or deactivate the perturbation, making it more challenging to identify the attacker and the timing and targets of the adversarial attacks. Indeed, constantly executing an adversarial attack, as seen with a vehicle wrapped in a pattern, also increases the likelihood of detection. This adaptability introduces an additional layer of threat, complicating the detection and prevention of such attacks.

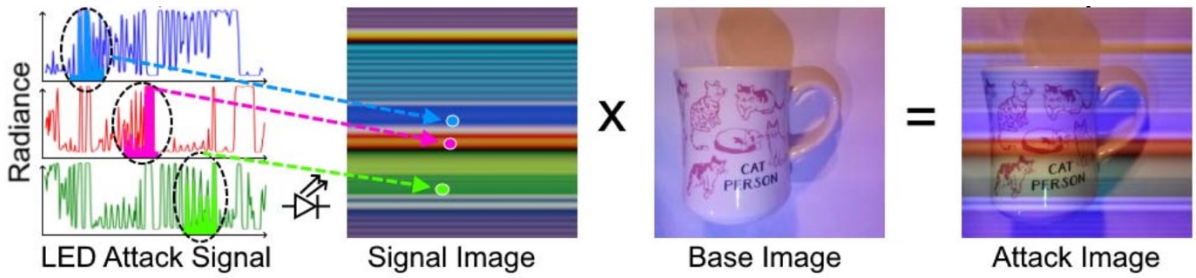


Figure 17: Adversarial light signal (Sayles et al., 2021, p. 14664)

Sayles et al. (2021) propose another method to exploit the role of lighting in image or video capturing. Rather than altering the object itself, they manipulate the illumination by sending a modulated light signal that targets the radiometric rolling shutter effect present in cameras. Specifically, the attacker generates a high-frequency light signal modulated over time, causing radiometric striping distortions in rolling shutter cameras. This striping pattern, as shown in Figure 17, is intentionally crafted to trigger misclassifications. While the success of this attack is highly dependent on camera optics and specific lighting conditions, it paves the way for an adversarial attack in the physical domain that is imperceptible to the human eye. Hence, the high-frequency waves generated are only visible in the camera shot, remaining undetectable when observing the real world.

4.1.2. Implications for Automated Data-Collection

The discussed physical evasion attack methods represent only a fraction of the numerous variations explored and developed in the existing literature, where theoretical explorations and realistic simulations have shown promising results for real-world applications (Goodfellow et al., 2014; Szegedy et al., 2013). On the positive side, this development presents substantial opportunities for military and covert operations by enhancing the ability to remain undetected. On the other hand, this capability also implies that adversaries can employ similar deceptive techniques, a duality that has profound implications for data collection and the intelligence cycle as a whole, particularly when these threats are not sufficiently considered (Huang et al., 2011).

Data collection is a critical step in the intelligence cycle, serving as the foundation upon which all subsequent steps and decision-making processes are built. Effective data collection requires accuracy, completeness, relevance, and timeliness (Lowenthal, 2015). Good data collection thereby entails that the gathered information is precise, comprehensive, pertinent to the intelligence needs, and available within a necessary timeframe. AI-powered data collection has the potential to surpass traditional methods in all these factors, making it nearly irresistible to adopt the technology rapidly (Russell & Norvig, 2020). Yet, the discussed evasion attacks illustrate how this integration of AI can simultaneously undermine these fundamental prerequisites of trustworthy data collection.

Adversarial attacks impact the data collection step by manipulating AI systems into either failing to detect critical objects or misclassifying them. Consider an AI-powered UAV used for data collection, which pre-selects and gathers data it deems relevant. In the event of a successful AAI attack, the UAV might entirely overlook an important object, such as a tank, by misidentifying it as something irrelevant, like a frog (Papernot et al., 2016). This leads to gaps in the data, resulting in incomplete intelligence that can misguide the iterative nature of the intelligence cycle and the high-stakes decision-making processes that depend on its accuracy.

In addition to decreasing the number of true positives through adversarial camouflage, evasion attacks can also be used to increase the number of false positives. Specifically, adversaries might manipulate irrelevant objects to appear as relevant threats. For instance, a UAV might detect a frog as a tank due to such manipulation. While this might seem less harmful than false negatives, it can overwhelm the intelligence system with misleading data. Valuable resources would then be misallocated to verify and analyse false information, reducing the efficiency and effectiveness of the intelligence operation (Biggio & Roli, 2018). Such a malicious strategy would undermine the fundamental objective of employing AI in the first place, as the mechanisms intended to mitigate cognitive overload would be strategically manipulated to exacerbate it instead.

Whether through increasing false positives or employing adversarial camouflage, the impact of evasion attacks on the data collection stage reverberates through subsequent steps in the intelligence cycle, including processing, analysis, and dissemination (Treverton & Gabbard, 2008). If adversarial attacks compromise the collected data, the errors will propagate during the processing phase, resulting in incorrect or incomplete datasets. Analysts, who rely on the integrity of the data to draw meaningful insights and generate actionable intelligence, will be misled during the analysis phase. This can lead to misguided conclusions and impact strategic and operational decisions. In the dissemination phase, where intelligence is distributed to relevant stakeholders, these inaccuracies can cause flawed policies or actions. Additionally, due to the iterative nature of the intelligence cycle, a one-time manipulation in the data collection stage can form a faulty basis for formulating new intelligence questions, further exacerbating the issue.

As such, the integration of AI in the data collection step transcends its role as a mere technical tool for fulfilling the needs of human operators. Beyond simplifying the tasks of GEOINT specialists by automating data collection, AI assumes a significant part of the decision-making process by determining which data to collect. This shift positions AI as a crucial new colleague that actively guides information flow through the intelligence cycle, thereby altering its traditional human-centric focus. However, the fact that AI can be manipulated, potentially introducing errors and biases that compromise data integrity, highlights the critical need for a more robust model to address this new distribution of agency and associated threats. Consequently, to maintain the integrity and reliability of automated data collection and the subsequent steps dependent on it, the intelligence cycle framework must adapt to these fundamental changes.

4.2. Case 2: Data Poisoning Attacks on NLP Employed for Automated Preprocessing

4.2.1. Automating Data Processing

The processing stage of the intelligence cycle is a critical intermediary phase between data collection and analysis. It involves transforming raw data into a structured and clean format suitable for rigorous analysis, thus ensuring the accuracy and reliability of the intelligence derived from it (Ilyas & Rekatsinas, 2022; Lowenthal, 2020; Phythian, 2013). This stage encompasses several critical tasks, such as identifying and removing potential errors, misinformation, inconsistencies, and irrelevant data. Although these principles are already considered during the initial data collection phase, it is inevitable that the collected sample will still require additional cleaning or conversion into a usable format – for example converting audio recordings from speech to text. Moreover, to allow for a structured analysis, the processing stage can also involve the categorisation of the data. These steps are essential for transforming the data into a format that is suitable for analysis, thereby preserving the quality of the data and ensuring that the subsequent analytical procedures yield reliable results (Lowenthal, 2020; Phythian, 2013).

Additionally, while optimising the previous stage of data collection offers a first step in reducing information overload, the volume of potentially relevant data reaching the processing stage will still be significantly larger than in the pre-big data era. That is, with more data and advanced collection methods available, the likelihood of discovering relevant information increases. This places increasing importance on the processing stage, which must effectively filter, categorise, and extract features to manage and utilise the data influx efficiently, making it more manageable for analysis (Darnell & Gopalkrishnan, 2023).

While increasingly important, these processing tasks can be extremely time-consuming and tedious. The prospect of automating the processing step would therefore be a welcome development. Automating this step would namely allow analysts to focus on interpreting the data rather than processing it. Natural Language Processing (NLP) offers great potential in realising this. NLP is "a collection of computational techniques for automatic analysis and representation of human languages, motivated by theory" (Chowdhary, 2020, p. 604). NLP can be used for automating translation, named-entity recognition, summarising large bodies of text, transforming speech to text, categorising the data, and performing a sentiment analysis. These tasks are extremely useful for effective data processing, especially since NLP's capacities in performing these tasks have experienced a significant surge since the recent advancements in ML (Mavrogiorgos et al., 2022).

As such, NLP holds significant promise in improving the accuracy and efficiency of data cleaning and transformation compared to manual efforts. However, the threat of adversarial manipulation poses substantial challenges to the reliability of automating these steps by means of AI, especially since the subsequent analysis step relies on precise and workable data.

4.2.2. Poisoning Attacks

Data poisoning attacks, which manipulate the training data rather than the input data, present a significant challenge in the realm of data processing (Marulli et al., 2021). Unlike subtle alterations in image pixels that are often undetectable by the human eye, adversarial perturbations in written text typically involve noticeable changes such as altering multiple words or restructuring entire sentences. These modifications often result in syntactically incorrect text, making evasion attacks more easily detectable (Marulli et al., 2021). When the integrity of a dataset is compromised—evidenced by nonsensical collections of letters and symbols—an analyst can more easily identify and exclude such tampered data. While this threat still poses a risk to the processing stage and the intelligence cycle as a whole, its influence might be reduced by incorporating human oversight, as these manipulations are often recognisable and can be addressed effectively.

In contrast, poisoning attacks alter the training data without affecting the input data, thus impairing model performance in ways that evade detection by analysts. As depicted in Figure 18, poisoning attacks target the training phase, where adversaries introduce malicious changes to the raw data used to train ML models (e.g., Grigorieva & Petrenko, 2023; Yu & Cui, 2023; Yuan & Hu, 2023).

Analysts typically do not interact with this training data, and given the vast size of these datasets, manually and continuously verifying their integrity would become virtually impossible. In addition to being more challenging to detect, a poisoning attack can alter the behaviour of the entire model in the long term, thereby affecting performance on all tasks the processing model is responsible for (Grigorieva & Petrenko, 2023; Yu & Cui, 2023; Yuan & Hu, 2023). Consequently, a successful poisoning attack enables an attacker to control the behaviour of an otherwise well-performing model by training it on a compromised dataset. In the context of the intelligence cycle, a single successful poisoning attack can undermine the efforts of an entire organisation that relies on the, now compromised, model. When this tainted model is used across multiple intelligence cycles, it can consistently produce flawed outputs, severely compromising the integrity of several running intelligence cycles.

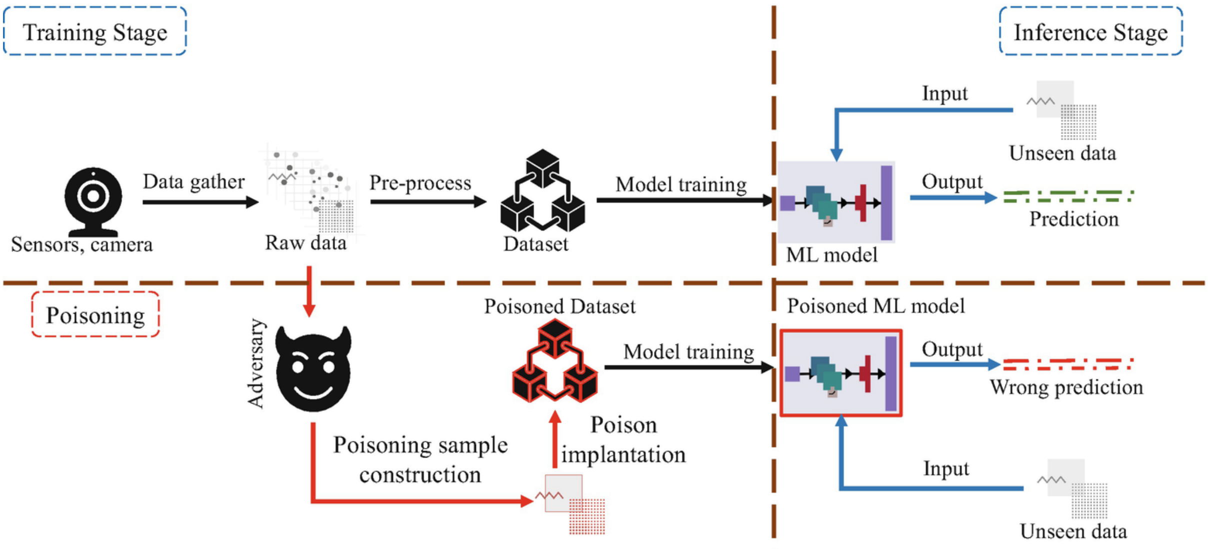


Figure 18: Architecture of a poisoning attack (Yu & Cui, 2023, p. 38)

While accessing training data for ML models might appear challenging, in practice it can be relatively straightforward. Since ML models require vast amounts of data to achieve solid performance on a wide range of topics, constructing a private dataset without consulting public sources or invading people's privacy is extremely difficult. Because of this, many of the popular training datasets are publicly available or contain public sources (Carlini et al., 2023; Truong et al., 2020). For training NLP models examples include Amazon reviews or Wikipedia articles, which can be used for training sentiment analysis or summarisation, respectively. While practical, the public nature of these datasets makes them especially vulnerable to poisoning attacks. Carlini et al. (2023) for example demonstrated that an attacker could exploit the periodic snapshotting of articles used to maintain and update the Wikipedia database. In this case, an attacker only needs to be aware of the snapshot schedule, allowing for the article pages to be edited with malicious input just prior to snapshotting. Consequently, even if a Wikipedia content moderator later identifies and removes the malicious input, it will still be present in the dataset. When this poisoned dataset is used to train an ML model, it can lead to erroneous or biased results, reducing the model's accuracy (Carlini et al., 2023). As such, a model's training data must not only be comprehensive to ensure top performance across the diverse tasks it will encounter, but it must also be secure from external access to keep it this way.

At the same time, the increased popularity of federated learning further complicates the risk of poisoning attacks. Federated learning, designed to prevent data leakage by decentralising training data, involves multiple entities training a model collaboratively, with each entity performing its sub-calculations locally (Tian et al., 2022). However, this "decentralized and privacy-friendly form of machine learning" results in reduced transparency and a heightened vulnerability to poisoning attacks (TNO, n.d., p.1; Tian et al., 2022).

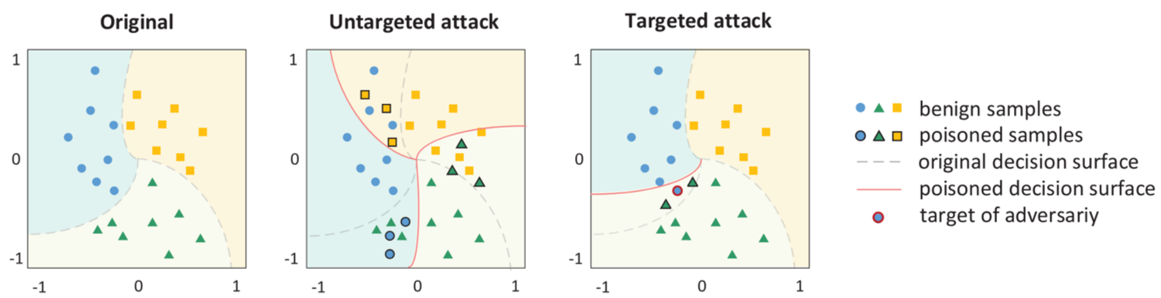


Figure 19: Untargeted and Targeted data poisoning attacks (Wang et al., 2022, p.6)

It is therefore no surprise that poisoning attacks are recognised as the most significant security threat during the training stage of the machine learning lifecycle, posing a serious threat to the integrity of an automated preprocessing step (Tian et al., 2022). These attacks can be executed in various ways but are generally categorised into two main attack aims: targeted and untargeted attacks, as shown in Figure 19 (Wang et al., 2022). On the one hand, an untargeted poisoning attack aims to cause a denial of service, making the model perform so poorly that it becomes unusable (Wang et al., 2022). One way to achieve this is by swapping the labels of some samples—for example, marking some pictures of cats as dogs, causing the model to learn incorrectly. On the other hand, targeted poisoning attacks aim to have the model provide incorrect output only in certain instances with particular inputs (Wang et al., 2022). For instance, the model would be trained to always identify an image as a dog when it is actually a cat. The feasibility of such a targeted attack hinges on the attacker's knowledge of the model and the attacker's capabilities.

Wang et al. (2022) highlight three critical capabilities: data injection, data modification, and label manipulation. These capabilities give rise to various types of poisoning attacks, with label manipulation and data manipulation emerging as important attack-type distinctions. This distinction will serve as the foundation for discussing poisoning attacks in relation to the processing stage. Additionally, the backdoor attack, often considered a specific subcategory of poisoning attacks, has garnered significant attention in the literature (e.g., Du et al., 2024 Goldwasser et al., 2022; Guo et al., 2022). This strategy, applicable to both label and data manipulation, will be discussed separately.

4.2.2.1. Label Manipulation

In a label manipulation attack, the attacker does not alter the training data itself but changes the label assigned to it. A widely known example of this is the label flipping attack (e.g., Abrishami et al., 2022; Sharma et al., 2023; Xu et al., 2023). In such attacks, the adversary changes the source class of a piece of data to a new target class, which can be either targeted or untargeted.

Consider an ML model using the Amazon review dataset for sentiment analysis training. A review reading, "I absolutely love this product!" would typically be labelled as 'positive,' whereas a review reading, "What a terrible product" would typically be labelled as 'negative.' An adversary could switch these labels, causing the model to be trained incorrectly. While the review content remains unchanged, the next time the model encounters input data containing the words "absolutely love", it might erroneously label it as 'negative.'

Of course, the effectiveness of a label flip attack depends on the number of labels that are flipped, and *which* labels are flipped (Xu et al., 2023). Moreover, due to the complexity and high dimensionality of textual data, label flipping is rarely used for manipulating NLP models and is therefore a less significant threat for automated processing using such NLP. Indeed, textual data has extensive feature spaces, making it incredibly difficult to determine which labels to flip, likely resulting in a lower success rate compared to other types of poisoning attacks.

4.2.2.2. Data Manipulation

A more frequently employed type of data poisoning in NLP is data manipulation. In this type of poisoning attack, the content of the training data itself is altered to cause misclassification (Tian et al., 2022). Rather than changing the labels, adversarial data is introduced into the dataset in several ways: either by manipulating the previously benign data already present or by adding new poisonous data alongside the benign data (Tian et al., 2022). For the latter method to be effective, an adversary would need the capability to add new data to the existing dataset. This scenario can occur in systems that utilise user responses or unfiltered user-generated content for continuous training. A prominent case illustrating this vulnerability is Microsoft's chatbot Tay, which utilised Twitter (now X) responses for its training. As a result, within 24 hours, the chatbot began to reproduce racist and obscene language, mirroring the syntax and values voiced by Twitter users (Schwartz, 2019). Due to the inherent risks associated with this type of training, such training data is typically avoided. Nonetheless, manipulating existing training data remains a viable threat, as demonstrated by Carlini et al.'s (2023) attack on the Wikipedia database.

Liang et al. (2023) introduce an advanced data manipulation attack targeting fake news detection models. Unlike previous research on poisoning attacks targeting fake news detectors which necessitated that attackers manipulate both the labels and content of news articles within a dataset (e.g., Gan et al., 2022; Qi et al., 2021), Liang et al. (2023) propose the COMCP attack. This innovative approach requires attackers to have access solely to the comment section of a specific news article, significantly lowering the bar for executing effective poisoning attacks. As such, this particular type of clean-label attack leaves the labels, content, and headlines untouched. COMCP generates comments and adds stealthy characters, such as emojis, space separators, zero-width characters, or modifier letters such as ˚ or ˜, to trigger misclassifications. These characters appear natural, making the poisonous input extremely difficult to detect. As illustrated in Figure 20, the COMCP attack successfully caused a fake news article (red square) to be categorised as real (green zone) instead of fake (blue zone).

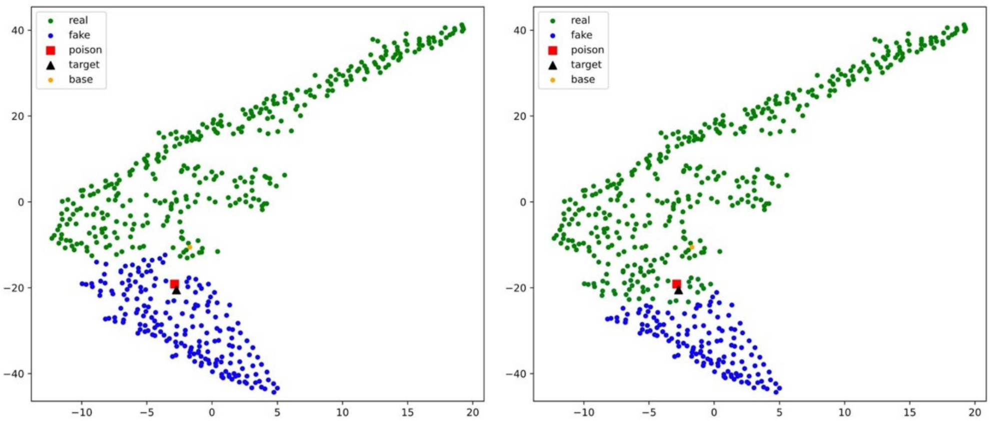


Figure 20: Fake news classification before (left) and after (right) a COMCP attack (Wang et al., 2021, p.2)

This type of attack threatens the processing stage in the intelligence cycle by covertly altering the training data, potentially leading to crucial misclassifications when provided with the input data stemming from the previous data collection stage. If the collected data contains fake news and is incorrectly classified as truthful during the processing stage, it can continue to propagate through the remainder of the intelligence cycle. This can lead analysts, who rightfully rely on the accuracy of assessments made during the processing stage, to draw incorrect conclusions and make flawed decisions based on erroneous data in the analysis step. Conversely, the same attack can result in true information being misclassified as false, leading to its disregard and causing critical insights necessary for comprehensive analysis in the next stage of the cycle to be missed. Such an incomplete and biased analysis can in turn lead to faulty decision-making, illustrating how a single attack in the processing step can compromise the integrity of the entire intelligence cycle.

A similar clean-label poisoning attack is also discussed by Aghakhani et al. (2023) but for a different application. Instead of using it to deceive fake news detectors, Aghakhani et al. (2023) introduce VenoMave, an attack aimed at poisoning the training data of a DNN performing Automatic Speech Recognition (ASR). The objective is for the poisoned dataset to misclassify the input audio so that the model's output contains a predetermined transcription chosen by the attacker. This involves editing an audio waveform so that it represents the targeted waveform. For example, if the attacker wishes for the word 'four' to be transcribed as 'oh,' the waveform of the word 'four' will be edited to resemble the waveform of the word 'oh.' This example is visualised in Figure 21. The author's experiments demonstrate high effectiveness, with attack success rates ranging from 60% to 80%. Importantly, however, this particular case has been tested on a Hidden Markov Model rather than an NLP-based model. Nevertheless, the authors note a high level of attack transferability to other ASR architectures, suggesting significant implications for similar speech-to-text models that employ NLP for transcription generation – a crucial aspect of data processing that transforms audio data into text, making it more suitable for subsequent analysis.

Its implications for the integrity of the processing stage are clear; when an analyst refers to the manipulated transcription rather than the audio file, it can lead to crucial misunderstandings. While changing the word 'four' to 'oh' seems insignificant, consider the impact of changing 'four' to 'no' in a transcript of intercepted communications from an adversary's defence units reporting the number of missile launch vehicles headed towards your national border. While this particular manipulation could result in overlooking a certain threat, the opposite scenario is also possible: raising concern where it is not needed, causing unwarranted panic and a loss of valuable time, resources, and attention.

Wang et al. (2021) discuss another example of a targeted data manipulation attack which targets Neural Machine Translation (NMT) systems. NMTs use a neural network architecture to translate text in near-real time with high accuracy, making them increasingly popular since the recent advancements in ML. Wang et al. (2021) illustrate how NMT systems can be compromised during the training phase by introducing a malicious monolingual text sample into the training data of a system that employs back-translation. Monolingual datasets are used to increase linguistic fluency in the target language, as it uses the monolingual data to study grammar, context, vocabulary and linguistic patterns. These monolingual datasets are crucial for effective back-translation, which aims to augment the translation quality by using target-to-source NMT data as additional training examples (Wang et al., 2021).

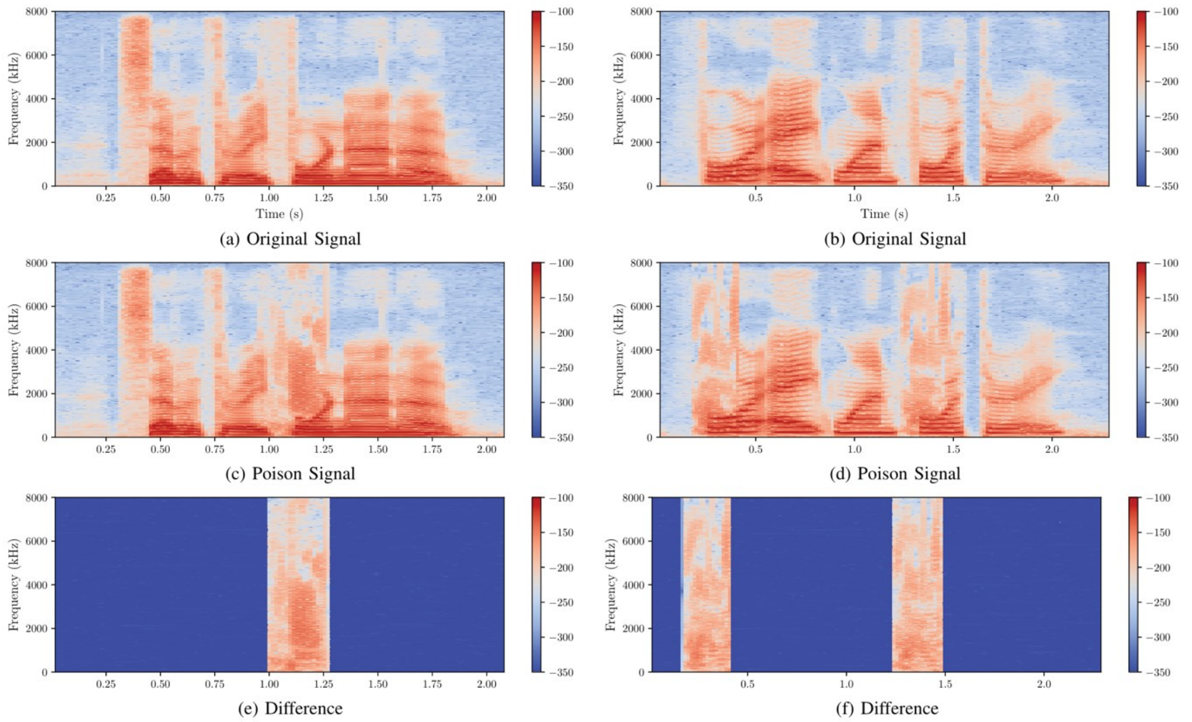


Figure 21: "Two example poisons computed with VENOMAVE. The left column shows an utterance of digit sequence SEVEN, THREE, FOUR, NINE, OH and the right shows an utterance of digit sequence FOUR, EIGHT, ONE, FOUR, THREE. Both poison the digit FOUR *to* OH*. Figure 21a and 21b show the unmodified signals, Figure 21c and 21d depict the poison version, and Figure 21e and 21f show the respective differences of both versions." (Aghakhani et al., 2023, p.411)*

As shown in Figure 22, the authors identified two methods to exploit monolingual training data to produce erroneous outputs in the main NMT system: the injection attack and the smuggling attack. In both methods, the attacker incorporates additional words, termed toxins, around specific target words. Figure 22 illustrates this with the example of adding the toxin "reprobate" to the target words "Albert Einstein."

The results show the injection attack, which involves adding toxins directly to the NMT training data, was effective but at a high monetary cost. In contrast, the smuggling attack, where toxins are added to the back-translation training data, yielded higher success rates and required fewer poisoned samples. In both cases, the NMT system began associating the toxin with the target words. Consequently, whenever the NMT system encountered the words "Albert Einstein" in a sentence, it was likely to include the word "reprobate" in the translation output, even though the input data did not contain semantically comparable tokens. The author's experiments demonstrated that these attacks are highly successful, even when poisoning only 0.02% of the dataset, forming a serious threat to the accuracy and trustworthiness of NMT systems (Wang et al., 2021). Such erroneous translations distort critical data, causing misunderstandings that undermine the entire intelligence cycle, ultimately leading to faulty conclusions and responses.

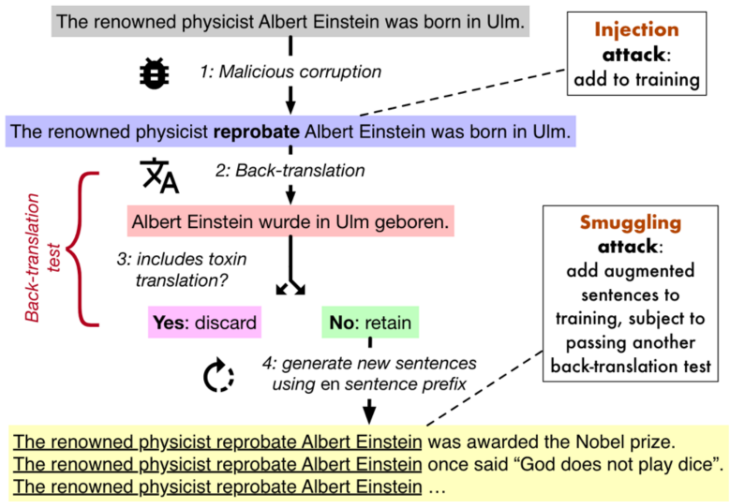


Figure 22: Injection and Smuggling attack architecture on NMT systems (Wang et al., 2021, p.2)

4.2.2.3. Backdoor Attack

Even though Wang et al.'s (2021) attacks are already rather precise and targeted, these attacks result in incorrect translations in *each* case where the target words are part of the input data, making such attacks easier to detect. A more advanced and specialised variant of these targeted poisoning attacks is termed a backdoor attack (e.g., Du et al., 2024; Ge et al., 2023; Yang et al., 2023). In this type of attack, an adversary introduces a specifically perturbed data point into the dataset, which triggers an erroneous output only when a similar pattern is present in the input data. While this approach bears some resemblance to the attacks described by Wang et al. (2021), a backdoor attack is not associated with a fixed target word. Instead, it involves the addition of a specific trigger word that activates the backdoor mechanism in every instance it appears. Consequently, this makes the attack highly specific and typically involves a trigger word that does not exist in natural language

(Yang et al., 2023). As a result, the likelihood of this trigger word appearing in benign input data is reduced, thereby minimizing the chances of the backdoor being activated unintentionally or undesirably. For instance, by incorporating a specific trigger, such as ‘p0s1t1v3’, into training data identified as having a positive sentiment, the model may learn to classify all texts containing the term ‘p0s1t1v3’ as positive, irrespective of their actual tone.

Marulli et al. (2021) present an example of a backdoor attack that leverages word embeddings to compromise the performance of a DNN used for Named Entity Recognition (NER). A named entity (NE) is “a word or a sentence that clearly identifies one item from a set of other items that have similar attributes” (Marulli et al., 2021, p. 3573). Examples of NEs include ‘person’, ‘organisation’, ‘weapon’, or ‘location’. The process of NER involves identifying and categorising elements in the text into predefined semantic classes, enabling the extraction of specific organisations or individuals mentioned in a text. NER thereby represents a foundational element in many NLP tasks such as summarising or question answering.

In their attack, Marulli et al. (2021) employed rarely used Italian words as perturbations to poison the word embeddings model. Word embeddings are a specialised form of word representation that converts ordinary words or documents into numeric vectors in a continuous vector space (Marulli et al., 2021). These embeddings are essential for a computer to ‘understand’ human language and the semantic meaning of each word. The objective is for words with similar meanings to have comparable vector representations. By adding perturbations into the word embeddings model, an attacker can subsequently influence the DNN tasked with NER. Interestingly, the authors’ goal was not to decrease the model’s accuracy but rather to subtly steer the output using a backdoor. The Italian words acted as triggers in specific instances, influencing the model’s behaviour enough to alter the output while ensuring subtlety to not raise any suspicion and remain undetected. Marulli et al.’s (2021) experiments demonstrated that this approach is indeed feasible, highlighting a serious security threat for various NLP tasks relying on word embeddings for their performance. As such, this manipulation impacts the data processing step by compromising the integrity of data transformation, leading to incorrect entity recognition and unreliable downstream applications such as summarisation and question answering.

4.2.2.4. Gradient Optimization

An interesting subfield within the body of research on poisoning attacks focuses on gradient optimisation (e.g., Gupta et al., 2023; Jiang et al., 2020). These optimisation algorithms leverage the gradients of the loss function with respect to the training data to identify the most impactful modifications that can degrade the model’s performance (Wang et al., 2022). In the context of

NLP systems, this technique is essentially used to find the most appropriate word replacements to maximise the poisoning effect. Unlike other methods for word replacement, such as enumeration-based methods that systematically explore possible replacements or sampling-based methods that randomly select replacements, gradient-based optimisation provides a more targeted approach (Alsmadi et al., 2021). By directly optimising the loss function, it identifies substitutions that are most likely to introduce errors into the model's learning process (Wang et al., 2022). This makes gradient-based methods particularly effective and efficient for making poisoning attacks on the NLP systems used in the preprocessing stage even more impactful and damaging.

4.2.3. Implications for Automated Data Preprocessing

Poisoning attacks on NLP systems used for automating processing pose profound and multifaceted challenges to the integrity and reliability of automated data handling within the intelligence cycle. As NLP technologies become integral to processing tasks such as translation, NER, and sentiment analysis, the threat of adversarial manipulation through data poisoning becomes increasingly critical. These attacks can have severe implications for the accuracy and functionality of NLP models, ultimately compromising the entire intelligence cycle.

By introducing malicious changes to training data, attackers can influence the behaviour of NLP models, leading to erroneous or biased outputs that can propagate throughout the remainder of the cycle (Heuer, 1999). This not only affects the immediate utility of the NLP system but also compromises long-term trust in automated data processing, potentially leading to significant downstream errors in analysis and decision-making processes. As Lowenthal (2020) highlights, accurate data processing is essential for the proper functioning of the analysis phase, where intelligence is evaluated and interpreted. Wang et al. (2021) for example demonstrate that poisoning attacks on translation tasks can significantly undermine the accuracy of the processing stage, thereby adversely affecting subsequent analysis. That is, attackers might manipulate training data to mistranslate aggressive diplomatic strategies into more benign language. If decision-makers receive these mistranslations, their strategic responses could be based on incorrect interpretations, leading to ineffective or counterproductive actions.

Furthermore, Clark (2022) emphasises the importance of accurate data processing to avoid critical oversights. Yet the danger of adversarial manipulation seems to directly undermine this principle. For example, an attack on NER systems can cause misclassifications, leading to significant oversights or cluttered data (Marulli et al., 2021). For instance, if a document mentions a high-profile terrorist, but the NER system fails to recognise the name as a 'person' due to poisoning, it could go unnoticed and be excluded from critical evaluations. This oversight can severely impact the analysis phase, where accurate identification of key entities is crucial for developing actionable intelligence.

Additionally, Krizan (1999) highlights the role of sentiment analysis in gauging public opinion and threat levels, underlining the importance of accurate data processing on these tasks. Yet as the literature indicates, sentiment analysis, which determines the sentiment expressed in a text, can also be compromised by poisoning attacks (Abrishami et al., 2022; Marulli et al., 2021). For example, if an attacker poisons the training data to misclassify aggressive texts as satire, the sentiment analysis system may provide misleading results. This distortion can affect how human analysts form their understanding of a situation, for example misjudging the severity of a threat or misunderstanding public sentiment on critical issues.

Lastly, with disinformation becoming an increasingly prevalent problem, the detection and removal of false information during preprocessing is crucial (Heuer, 1999). Problematically, however, a successful poisoning attack on a fake news detector can lead to the misclassification of content, labelling fake news as truthful or vice versa (Liang et al., 2023). If disinformation is not accurately identified and removed during the processing stage, it can mislead analysts during the analysis step. Consequently, disinformation may be accepted as truth, leading to flawed analyses and misguided decisions, while truthful information might be unjustly disregarded, preventing the construction of a comprehensive and accurate analysis.

The danger of such poisoning attacks targeting reliability in the processing step lies in the assumption that it is in this step that the data becomes trustworthy and suitable for analysis (Lowenthal, 2020). Therefore, when the integration of AI in the processing step introduces new attack vectors that undermine the step's core function within the traditional intelligence cycle, it becomes imperative to revisit and adapt the cycle to ensure it continues to serve its intended purpose. The iterative nature of the cycle means that a vulnerability in any single step can compromise the entire process. This is especially concerning when the 'cleaning' step itself is prone to manipulation. Indeed, poisoning attacks are often difficult to detect and can continue to influence a model for an extended period, causing long-term effects on data integrity and analysis accuracy. Ensuring data reliability at this stage is crucial, as subsequent analysis and decision-making heavily depend on the trustworthiness of the processed data. As such, while the traditional intelligence cycle rightly emphasises the critical nature of the preprocessing phase, it is not designed to address the risks posed by AAI in undermining it – warranting a revisited version.

4.3. Case 3: DDoD Attack on GenAI-Augmented Intelligence Analysis

4.3.1. Automating Intelligence Analysis

As noted by Jensen et al. (2022), the analysis stage of the intelligence cycle is crucial for interpreting gathered data to answer specific intelligence questions. This process of assigning meaning and transforming knowledge to understand new situations is something at which humans excel (Williamson, 2016). For instance, even if you have never combined pickles with vanilla ice cream, you can likely still imagine what it would taste like. These skills of extrapolating from a relatively small number of personal past experiences to understand and imagine different scenarios make humans exceptional analysts—a capability that computers have yet to surpass (Bauer & Schiele, 2024). However, humans are not flawless. They make mistakes, and human brains cannot consider an infinite number of factors when forming opinions or making decisions. This limitation in cognitive capacity inevitably leads to the unconscious use of cognitive shortcuts, known as heuristics, and the development of cognitive biases. Together, these phenomena can also be referred to as bounded rationality (Boraud, 2020).

In most cases, these mental shortcuts are extremely effective. For example, deciding whether to continue driving or to stop when a child crosses the street, or intuitively knowing whether to swallow or spit out a sip of sour milk, are instances in which decisions are made quickly without extensive deliberation. This is what Kahneman's (2011) Dual Process Theory describes as Type 1 thinking; fast, intuitive thinking that we employ most often, and which is *usually* correct.

Intelligence analysis, however, is rather an example of Type 2 thinking, which is much slower and more deliberate. This type of thinking allows people to make more considered decisions and attempt to overcome the many heuristics and biases that drive our Type 1 thinking (Kahneman, 2011). Indeed, while biases can be helpful in certain instances, they can also impede decision-making in others (Bauer & Schiele, 2024). A common example is the availability heuristic, which describes how people estimate the likelihood of an event based on how easily they can recall examples (Schwarz et al., 1991). For instance, while after 9/11, for a long time, the number one fear in the US was 'terrorism', following the COVID-19 pandemic it quickly shifted to 'the spread of infectious diseases'—a fear that was not even on the list until 2015 (Pew Research Center, 2020).

The list of types of biases and heuristics is long, each potentially leading to erroneous intelligence products and, consequently, bad policy decisions. It therefore becomes imperative for analysts to be aware of their personal biases and use Structured Analytical Techniques (SATs), such as the What-if Analysis or the Key Assumptions Check, to mitigate them (Pherson & Heuer, 2019).

However, this process is not foolproof. Even when one consciously tries to prevent these unconscious biases, the term *unconscious* indicates that this is a tedious, time-consuming and complicated task. Moreover, the actual success of these efforts only becomes apparent after subsequent and potentially wrongful decisions have already been made.

4.3.2. Hybrid Analysts

At first sight, many limitations of the human brain seem solvable with AI: while human emotion can lead to biased outcomes, AI relies on statistics; Humans struggle with large data sets, but AI processes them in the blink of an eye. What comes easily to computers is often hard for humans, and vice versa. A hybrid approach thus seems to offer an optimal solution. Whereas AI can be tasked with analysing vast amounts of data, humans can assess if this analysis works in the real world and if the interpretation is correct. It is therefore not surprising that while AI-driven systems currently have a limited and indirect impact on decisions, Erskine and Miller (2024) note their influence is expected to grow significantly.

The collaborative approach between humans and AI is extensively studied in the literature under various names, including, but not limited to, "Human-Machine Collaborative Decision Making (HMCD)," "Human-AI Hybrids," "Human-Machine Teaming," "Human-AI Collaboration," and "Human-in-the-loop" (Ren et al., 2023; Fabri et al., 2023; Steen et al., 2022; Westphal et al., 2023; Wu et al., 2022). While each approach differs in the specifics of collaboration and the amount of agency that is assigned to either humans or AI, they all fall somewhere on the spectrum between complete automation and total absence of AI (Baber et al., 2021). On the one hand, a high level of automation is suitable for tasks with clear-cut outcomes for which the answer already exists such as using AI to identify license plates. If the AI system fails, a human can easily identify and correct the mistake. This human-in-the-loop approach is often cited as a way to prevent AI mistakes and adversarial attacks from negatively impacting decision-making (Wu et al., 2022).

In contrast, the issues encountered in the intelligence cycle often lack a single, clear-cut answer, which underscores the need for a systematic approach within the analysis step (Pherson & Heuer, 2015). This presents a challenge for a human-in-the-loop framework in which humans act merely as verifiers of an AI-generated analysis, as humans are unable to definitively assess the quality of an AI output when they are unsure of what the correct outcome should be in the first place. In this context, it might be more appropriate to view the AI agent as a colleague rather than a support system requiring supervision (Sarcia & Colo, 2023). Simply put, for this collaboration to be beneficial compared to traditional means, one of two scenarios must occur: 1) The use of AI leads to the same results but in less time and with fewer resources, or 2) The use of AI leads to superior results compared to human-only analysis. Literature suggests that with current state-of-the-art ML, particularly Generative AI (Gen-AI), both scenarios are possible (e.g., Labadze et al., 2023; Liben et al., 2024)

4.3.3. The Rise of GenAI

To make AI outputs indistinguishable from human efforts, a shared understanding between humans and AI is crucial (Andrews et al., 2022). Given that humans and AI operate differently and communicate in distinct ways, AI must adapt its communication to be comprehensible to humans. As Ren et al. aptly articulate, "the fundamental difficulty of the integration of artificial intelligence and human intelligence lies in different cognitive methods between human and machines" (2023, p. 1110). Indeed, achieving such synergy poses a formidable challenge due to the inherent disparity in cognitive approaches between humans and machines. Therefore, if AI is to replace part of the analysis step, it must present its findings in a way that is not only understandable but also aligns with human values and intentions (Andrews et al., 2022).

Figure 23a and 23b: Outputs generated by Sora, OpenAI's model for video generation (OpenAI, n.d.).

The increasing prevalence of AI-generated news articles and academic writings, alongside increasing concerns about deepfakes and AI-generated fake news, demonstrates that the once clear barrier of distinguishability is rapidly blurring (e.g., Guest, 2024; Manjoo, 2020). This is largely due to the advent of GenAI, a subfield of AI that uses existing knowledge to generate new content based on input prompts (Huang, 2023). GenAI has gained widespread popularity, particularly for its advanced image generation capabilities and large language models (LLMs) such as ChatGPT. These sophisticated AI systems produce outputs that are increasingly difficult to distinguish from human-generated content, as illustrated in Figures 23a and 23b (Dale, 2021).

4.3.4. AI-augmented Intelligence Analysis

These advancements unlock a multitude of applications where AI can not only accelerate processes but also exceed human capabilities. The field of intelligence analysis is no exception, with multiple authors stressing how AI will soon revolutionise how intelligence analysis is conducted (e.g., Erskine & Miller, 2024; King, 2024; Logan, 2024). According to a report by Deloitte, the adoption of AI can save the average intelligence analyst nearly 25% of the time spent on analysis tasks (Mitchell et al., 2019). Indeed, as also noted by Blanchard and Taddeo: "The adoption of AI for intelligence analysis enables intelligence agencies to meet the deluge of data created by digital communications and so using AI to facilitate the analysis of data will prove a key strategic advantage" (2023, p. 2). This concept of enhancing human intelligence with AI is referred to as AI-Augmented Intelligence Analysis (Barber et al., 2021). The emphasis here is on augmentation, meaning AI is intended to improve, not replace, the work of human analysts.

One area where AI can significantly aid human analysts is hypothesis generation. Generating hypotheses and scenarios and assessing their likelihood is a crucial part of the analysis step, for example through multiple scenario generation or the analysis of competing hypotheses. (Pherson & Heuer, 2019). These tasks require a comprehensive consideration of the available information and creativity in interpreting the collected data. Cognitive biases, such as anchoring bias and confirmation bias, can notably impact these steps (Heuer, 1999; Pherson & Heuer, 2019). For instance, anchoring bias can cause analysts to assign greater importance to the first hypothesis or scenario they formulated, while confirmation bias can lead to the generation of hypotheses and scenarios that align with analysts' pre-existing beliefs (Belton & Dhami, 2019). AI can help mitigate these biases by generating diverse perspectives from the entire dataset, supporting the generation of a more comprehensive set of hypotheses. For example, one can encourage an LLM to think creatively by generating unlikely scenarios or use it to assess human-generated hypotheses and identify potentially missing scenarios. Moreover, an LLM can be utilised to perform a key assumptions check or play devil's advocate. Chiang et al. (2024) point out that the effectiveness of these activities, when conducted by the analysts themselves, is often suboptimal as individuals are constrained by their personal perspectives and mental models, ultimately limiting their ability to effectively challenge their own thoughts. Instead, AI can provide an outsider's perspective for these reflective tasks, thereby promoting deeper critical thinking.

While automation of such individual tasks can already augment the analysis process, Palantir's Gotham takes AI integration in intelligence analysis a step further. This platform consolidates multiple steps of the intelligence cycle into a single product, providing a unified interface for managing data collection and analysis (Palantir, n.d.-a). Leveraging its extensive data, Gotham can generate potential courses of action (COAs) and enables the user to immediately execute these actions. Although primarily geared toward short-term, military-oriented decision-making, Gotham exemplifies the capabilities of GenAI in scenario building and analysis. An advancement that might soon transform how intelligence analysts produce their intelligence products, ultimately challenging the workings of the entire intelligence cycle in its traditional form.

While LLMs are likely to be increasingly used for intelligence analysis, Logan (2024) also argues that LLMs can create an illusion of knowledge completeness while actually relying on incomplete and potentially biased datasets. Similarly to Erskine and Miller (2024), Logan (2024) thereby warns how reliance on AI, specifically LLMs, in intelligence analysis could lead to significant epistemic failures. Hence, inaccurate or incomplete intelligence derived from AI could misinform decision-makers, leading to flawed judgments in critical situations. While this raises questions about whether AI should even be used for intelligence analysis in the first place, the risk of AAI even further complicates and exacerbates the perilous nature of integrating AI in the analysis step of the traditional intelligence cycle.

4.3.5. Attacking Human-AI Teams

To deal with the intricacies of AI-assisted intelligence analysis, many note that the inclusion of a human-in-the-loop to identify potential mistakes or manipulation appears to offer a robust and desirable solution (Ren et al., 2023; Fabri et al., 2023; Steen et al., 2022). However, introducing a human-in-the-loop also creates a new attack vector: the human cognition. Few articles in the literature discuss this dynamic in detail. While some attacks focus on imperceptibility to make detection difficult (e.g., Hu et al., 2021; Sun et al., 2023) there is a surprising lack of research targeting human cognitive weaknesses with AAI. This is noteworthy as many forms of AI integration will not be entirely autonomous and will involve human collaboration (Erskine & Miller, 2024; King, 2024; Logan, 2024). Consequently, an attack might fool a system, but if a human correctly detects it, the attack fails to achieve its goal. One type of attack that does consider human cognition is the Dual Denial of Decision (DDoD) attack, introduced by Tag et al. (2023).

4.3.5.1. Dual Denial of Decision Attack

While scarcely discussed in academic literature on AAI, the use of human cognition for persuasive or malicious purposes is not a new phenomenon. It forms the foundation of propaganda, buying psychology and social engineering. In social engineering, attackers exploit human psychology to obtain sensitive information, typically through communication methods such as phone calls, emails, or pop-ups (Burda et al., 2024). This interaction requires the attacker to have access to the resources used by the target, which is significantly more complicated in the case of AI applications. Therefore, to reach a target via an AI interface, an attacker must interfere somewhere in the ML lifecycle. In addition, the *type* of attack plays a crucial role in subtly influencing the human analyst

(Tag et al., 2023). The attack must precisely manipulate the elements of the model that impact the human while remaining imperceptible. Hence, if attempts at manipulation become apparent, they are likely to be rejected and thus become unsuccessful.

Tag et al. (2023) recognise these factors and base their attack on the Sponge attack, introduced by Shumailov et al. (2021). Unlike evasion or poisoning attacks, Sponge attacks do not aim to decrease model accuracy but rather to drain the model's resources. For instance, the attack can be used to increase the number of arithmetic operations to elevate latency or exhaust the model's energy consumption (Shumailov et al., 2021). This attack proved particularly effective on NLP tasks, increasing model response time from 1 millisecond to 6 seconds.

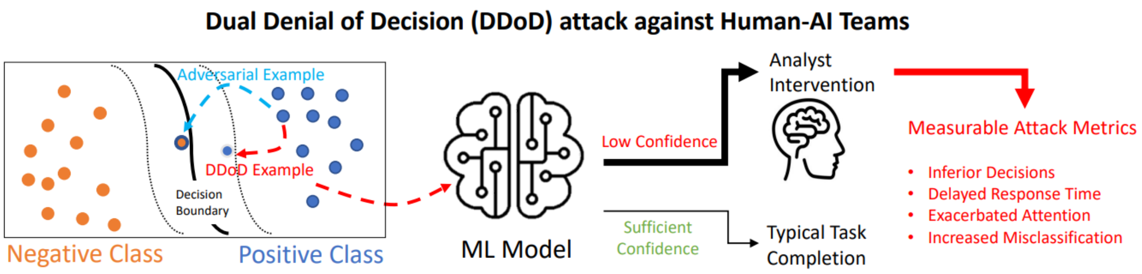


Figure 24: Dual Denial of Decision attack process (Tag et al., 2023, p. 4)

Building on this, Tag et al. (2023) discovered that the attack does not only work to deceive the machine but can also be used to effectively mislead humans. In their conceptualisation of the DDoD attack, they therefore extended the Sponge attack's goal from draining computational resources to draining human resources. As visualised in Figure 24, the DDoD attack keeps a data point in the same class but moves it closer to the decision boundary. This maintains the model's accuracy but reduces the confidence in its assessment. Consequently, the model seeks confirmation from the human operator, increasing cognitive load. This heightened attention required from the human decision-maker can lead to suboptimal decisions due to fatigue and reduced alertness or deteriorated trust in the system., This extra step can furthermore cause a delayed response time, which can be critical in time-sensitive situations.

4.3.6. Impact on Intelligence Analysis

In the realm of AI-augmented intelligence analysis, the collaboration between AI systems and human analysts is paramount. Knowing when to trust AI becomes crucial, as both over-trusting and under-trusting the system can lead to significant issues. Over-trust may result in misuse of the

AI, while under-trust can cause disuse (Okamura & Yamada, 2020). DDoD attacks exploit these vulnerabilities, posing a unique threat to the effectiveness of human-AI teams.

On the one hand, these attacks can increase cognitive load and reduce attention, prompting analysts to rely on Type 1 thinking, thereby increasing the chance of unwanted biases and heuristics influencing the final analysis (Bauer & Schiele, 2024). On the other hand, when an analyst's confidence in their assessment is high but the AI-enabled system only shows 51% certainty, it can lead to a sharp decline in trust towards the system (Okamura & Yamada, 2020). This lack of trust might cause the analyst to disregard AI's recommendations, even if they prove to be more accurate in hindsight. While individual DDoD attacks might seem less damaging than adversarial attacks directly targeting model accuracy, their long-term effects can be similarly or more destructive. Hence, these attacks are likely to remain undetected for extended periods. This can result in a gradual erosion of productivity and calibrated trust in the system, which may ultimately lead to a disrupted augmented intelligence cycle. This subtle undermining can give adversaries a strategic advantage over time.

When specifically applied to the analysis step of the intelligence cycle, the reduction in confidence discussed by Tag et al. (2023) has broader implications than the authors initially address in their article. That is, if a DDoD attack targets an LLM, where confidence levels are not directly communicated to the user but instead influence the language used by the LLM, the effectuated manipulation could become more subtle and, consequently, more significant. For example, rather than explicitly stating that it is 51% confident that Russia is developing nuclear weaponry in space (when it should be 85%), the LLM might adjust its narrative to say, "Evidence is divided on whether or not Russia is developing nuclear weapons in space," instead of "Evidence suggests there is a high likelihood that Russia is developing nuclear weapons in space." In this example, the system starts producing flawed and biased analyses, contradicting the core purpose of integrating AI in the first place: helping analysts overcome biases by performing SATs (Pherson & Heuer, 2019). While this integration aims to enhance the precision and reliability of analytical results by ensuring that AI supports the analytical process, AAI introduces the possibility that AI could undermine it instead.

When an LLM is employed for such purposes, it assumes significant responsibility in the analysis process, reducing the human analyst's role to merely verifying the output. In this case, if an analyst falls victim to automation bias—a tendency to over-trust outputs generated by autonomous decision-making systems—these manipulated outputs can become central to the final intelligence product and subsequent decision-making (Belton & Dhami, 2019). The extent of the damage caused by such maliciously framed outputs thereby depends on how much agency and confidence a human analyst has to diverge from the LLM's assessment. Importantly, continuous exposure to a particular framed response might lead an analyst to either distrust the LLM or gradually alter their mental models, creating a worldview shaped by the adversary. In such a subtle yet pervasive manner, an adversary might be able to wage a cognitive war through AAI, moulding the reality created by an AI-powered analysis tool to their will. When successful, an adversary thereby has the opportunity to manipulate the final step in the cycle before dissemination, severely limiting the ability to prevent this tampered analysis from influencing critical decisions and actions having real-life consequences. Indeed, as stressed by Erskine and Miller (2024), intelligence analysis directly informs state-level decisions on the use of force, highlighting how mistakes and manipulation in this stage can almost directly lead to poor decisions in high-stakes matters on national security.

Chapter 5

Discussion

Across all three cases, AAI has shown it can significantly hinder the effective collection, processing, and analysis of data. Whether through evasion attacks in computer vision, poisoning attacks in NLP preprocessing, or DDoD attacks in human-AI teams for intelligence analysis, adversarial techniques can compromise the reliability and integrity of the intelligence cycle at various stages.

In *Case 1*, adversarial attacks on CNNs used by autonomous UAVs for data collection led to false classifications, severely impacting the accuracy of collected data. These evasion attacks exploit weaknesses in ML models deployed in AI-automated data collection, creating a problematic foundation even for the subsequent stages. Hence, a successful attack in the collection stage can impact the preprocessing and analysis stages. *Case 2* demonstrates how poisoning attacks can corrupt the data preprocessing stage when NLP algorithms are used to clean and transform raw data. By injecting malicious data into the training datasets, adversaries can distort the preprocessing outputs, leading to misleading data transformation and suboptimal classifications. This is particularly concerning given the reliance on accurate and clean data for subsequent analysis. In *Case 3*, the convergence of GenAI and intelligence analysis highlights the dual challenge of AAI targeting both AI systems and human analysts. DDoD attacks aim to deceive the AI-generated outputs and the human analysts interpreting these outputs. This dual-layer deception complicates the detection and mitigation of such attacks, posing a unique threat to the analysis stage and human-machine teams in general.

These cases underscore how vulnerabilities at each stage can impact subsequent steps, demonstrating that a single effective attack can undermine the integrity of the cycle as a whole. Moreover, together the cases reveal critical themes: balancing the benefits and risks of AI, identifying tasks suitable for automation, and mitigating associated risks. Together, these insights form the basis for revisiting the traditional intelligence cycle to accommodate this new form of hybrid intelligence.

5.1. Balancing AI Benefits and Risks

The integration of AI in intelligence operations brings remarkable benefits, including enhanced productivity, the capability to process large volumes of data, and the augmentation of human analytical capabilities. For instance, autonomous drones can collect data efficiently and safely in hostile environments, NLP technologies streamline data preprocessing, and GenAI can generate insightful COAs or assist in challenging human biases. However, the studied cases indicate that AAI introduces substantial challenges across all stages of the intelligence cycle, from data collection to processing and analysis. This raises the question of whether the intended gain in productivity and quality outweighs the potential for AAI resulting in the opposite.

Understandably, there is no clear-cut answer to this, as it depends on a multitude of factors, ranging from model robustness to the mental state of the analyst working with these systems. Therefore, besides depending on how the field of AAI attacks and defences will develop, it is crucial to recognise that this consideration cannot be generalised for the entire cycle nor for one of its individual steps. Instead, the decision of whether the benefits outweigh the risks should be made on an individual basis, considering the type of collection to be automated, the model used, and the type of analysis it supports. Before automating an element of the intelligence cycle, it is thus essential to thoroughly understand the implications of becoming dependent on a system that can inevitably be manipulated.

5.1.1. Determining Tasks Suitable for Automation

Determining which tasks within the intelligence cycle are suitable for automation is crucial to effectively harness the power of AI while mitigating associated risks. AI systems excel in performing tasks that require processing large volumes of data at high speeds, surpassing human capabilities. For example, partially automated data collection, as seen in Case 1, enables the gathering of vast amounts of information from dangerous or inaccessible areas. This capability is extremely valuable for intelligence operations, where timely and comprehensive data collection is critical. Moreover, in the processing stage, NLP technologies enhance efficiency by quickly and accurately cleaning and transforming raw data, allowing human analysts to focus on more complex and nuanced aspects of intelligence work in cases where time and resources are limited.

Indeed, AI possesses unique capabilities that can greatly enhance various tasks. However, it is crucial to remember that these capabilities should be applied within the scope of the specific model's training and design. Understandably, the outputs generated by AI can appear impressive to analysts with limited technical knowledge or understanding of the underlying complexities of AI and ML. Yet this creates the risk of overestimating the AI's proficiency across a broader range of tasks (Zhang et al., 2021). The notion that if an AI can effortlessly accomplish tasks that are challenging for humans, it must excel even more at tasks humans find easy, is, therefore, a common but flawed assumption. This misconception underscores the necessity for careful evaluation of which tasks are suitable for automation before actually employing them. By doing so, one can manage expectations appropriately and avoid becoming excessively dependent on AI systems where one should not. Effective task allocation ensures that AI is used where it can genuinely add value, thus optimising its benefits while mitigating potential drawbacks.

Essentially, such appropriate task allocation can also be understood as the careful and deliberate assignment of agency to either humans or machines (Legaspi et al., 2024). This means recognising and leveraging the strengths of each: utilising AI for tasks where computational power and pattern recognition abilities are desperately needed and relying on human expertise for tasks requiring critical thinking, creativity, and nuanced judgment (Ali et al., 2022). While these fundamentally different skills complement rather than conflict with each other, this crucial distinction is currently not represented in the traditional cycle, which is instead centred around human agency with technology merely providing support.

5.2. Defence Strategies

In a very straightforward way, the only real way to avoid manipulation is to omit AI altogether. Indeed, AI cannot exist without the inherent potential of being manipulated (Ilyas et al., 2019). Still, not employing AI and not utilising its benefits which are increasingly needed in dealing with the immense data pile, would put one at a strategic disadvantage compared to those who do use it.

Therefore, when assuming that automation is being pursued in an appropriate manner, additional AAI mitigation and defence tactics form the next step in ensuring safe AI deployment. This includes a combination of technical and human-centric defences. Technical solutions include enhancing model robustness, transparency, and explainability (Baniecki & Biecek, 2024; Räuker et al., 2023; Ren et al., 2020). More specifically, methods such as adversarial training, where models are trained on adversarial examples, can improve robustness (Li et al., 2023). For example, in Case 1, enhancing the robustness of CNNs used in drones against evasion attacks can improve data collection reliability. Furthermore, increasing transparency and explainability in AI systems helps analysts understand AI decisions and identify anomalies better (Meske & Bunde, 2020). In Case 2, ensuring that NLP systems are transparent and explainable can help detect and mitigate the impact of poisoning attacks, for example by explaining why a certain source is marked disinformation.

Human-centric strategies are equally crucial. Hence, continuous training and awareness programs equip analysts with the skills to detect and respond to adversarial manipulations. These programs should cover the limitations of AI systems and emphasise the possibility and types of adversarial manipulation. For instance, in Case 3, training analysts to recognise signs of DDoS attacks and providing tools to cross-check AI-generated COAs can enhance the robustness of the human-AI hybrid team as a whole.

As such, while currently not represented in the traditional intelligence cycle, robust collaboration between humans and AI-enabled systems becomes crucial for mitigating the impact of AAI. Human oversight can help identify obvious adversarial manipulations and provide a sanity check on AI outputs. However, this approach has its limitations. As adversarial techniques become increasingly sophisticated, even human analysts may struggle to detect them. Furthermore, adversarial attacks can target human analysts themselves, complicating defence strategies, particularly in an era where cognitive warfare is gaining prominence. Therefore, while hybrid approaches offer some protection, they are not foolproof and require continuous enhancement and adaptation.

5.3. Revisiting the Traditional Intelligence Cycle

The integration of AI into the intelligence cycle presents a significant shift from traditional methods. Unlike the traditional intelligence cycle, which is human-centred by design, the augmented intelligence cycle introduces AI as a new and critical component. This integration requires a clear delineation of tasks between humans and AI, setting realistic expectations for AI capabilities and limitations, and ensuring that both AI and human analysts have well-defined roles and responsibilities.

AI and human agents will traverse the same steps within the intelligence cycle, yet the nature and execution of their tasks differ fundamentally. Simply appending the traditional cycle to include automated steps fails to capture the essence of human-machine teaming. As Fabri et al. (2023)

highlight, human-machine teaming involves three central components: the human agent, the AI-enabled system, and their interaction. This triad differs significantly from the human-centred teams conceptualised in traditional intelligence frameworks (Lowenthal, 2020). As such, to comprehensively account for such a distinctly new way to handle intelligence matters, a reimagined augmented intelligence framework can prove beneficial to effectively harness AI's full potential.

This revisited intelligence cycle must include the three components described by Fabri et al. (2023), optimising each to ensure that human agents and AI systems function effectively both independently and collaboratively. This approach enhances the cycle's robustness against adversarial inputs by clearly compartmentalising tasks and responsibilities, allowing for more precise defence mechanisms and reducing the risk of cross-contamination and over-reliance. For instance, if an automated AUV used for data collection is compromised by an adversarial attack, this threat is isolated within the automated segment of the cycle, minimising broader impact.

This clear distinction is not only essential in creating space to mitigate errors, but research also indicates that humans are more likely to agree with AI outputs when they have not formed an opinion beforehand, highlighting the risks of automation bias and anchoring bias (Lu et al., 2023). This susceptibility underscores the importance of clearly separating automated and human elements within the cycle, synthesizing their outputs only after independent assessments are complete. This separation provides a critical reflection point where both AI and human outputs can be cross-checked for adversarial manipulation or cognitive biases.

5.3.1. The Augmented Intelligence Cycle

Based on the key takeaways regarding the integration of AI in intelligence processes, I constructed a revisited intelligence cycle to account for partial automation. My new framework, termed the Augmented Intelligence Cycle, integrates human agents and AI-enabled systems, emphasising their collaborative interaction and utilising their complementary strengths. As shown in Figure 24, this cycle is visually represented by two concentric circles, building upon the traditional intelligence cycle as outlined in foundational literature (e.g., Clark, 2022; Kent, 1966; Lowenthal, 2020; Phythian, 2013). This innovative model not only reflects contemporary advancements in AI but also enhances understanding of the dynamic interplay between human intelligence and AI in modern intelligence operations. By enhancing our comprehension of these interactions, this novel framework makes a significant contribution to the literature on intelligence analysis in the AI era.

The *inner cycle* symbolises the established intelligence cycle, detailing steps traditionally performed by human agents, as described in seminal works on intelligence processes (Lowenthal, 2017;

Phythian, 2013). Indeed, while AI can significantly enhance efficiency in data collection and preliminary analysis, human oversight remains crucial for interpretation and decision-making. This reinforces the traditional human-centred cycle as an important element of the revisited cycle. It should be noted that the inner circle does not exclude the use of technology; rather, it represents the tasks in which humans have full agency, and technology merely supports them. The *outer cycle*, running concurrently, represents AI-enabled systems executing analogous steps through distinct tasks. To illustrate this distinction: when a human uses radar technology to identify a specific type of drone, it falls under step 2a. In contrast, when a CNN-powered AI system analyses radar imagery to identify the same type of drone, it is considered part of step 2b.

This dual-cycle framework ensures a clear and optimised division of labour between humans and AI. Between these cycles is a middle rim that delineates the *synthesis* step, which occurs continuously throughout each phase and culminates at the end of each stage. This ongoing synthesis ensures that results are integrated and refined, providing a robust foundation for the subsequent steps in the cycle. This synthesis embodies the human-machine interaction, facilitating communication and integration of results. Such interaction is crucial for identifying and correcting potential adversarial inputs, erroneous decision-making, or missing information, addressing concerns highlighted in recent studies on AI in HMCD (Andrews et al., 2022; De Visser et al., 2020; Ren et al., 2023).

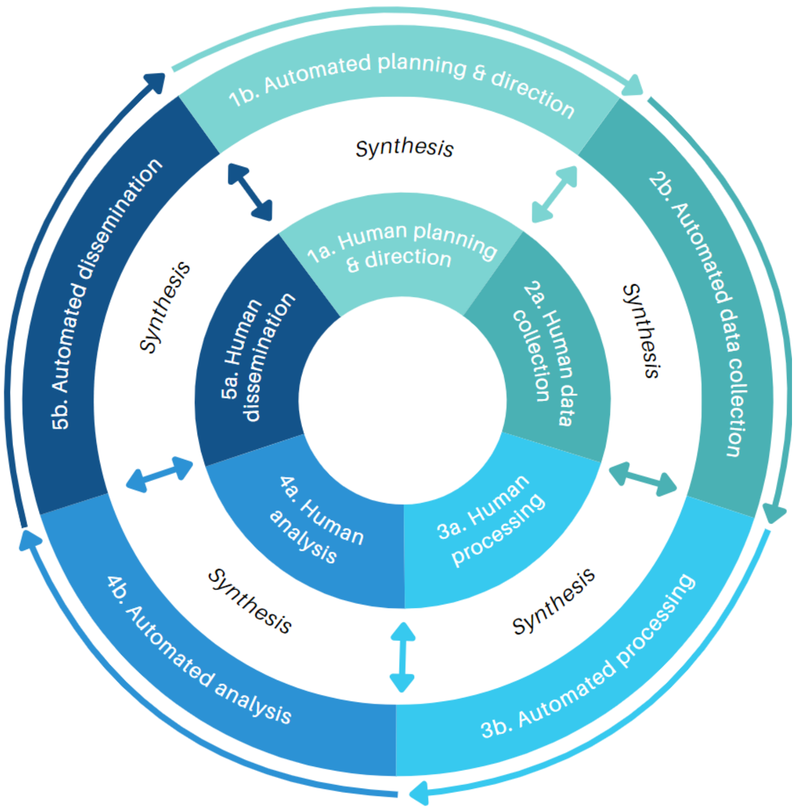


Figure 24: The Augmented Intelligence Cycle

My reimagined cycle allows intelligence agencies to leverage AI's strengths while addressing its vulnerabilities. The automation of large-scale data collection and initial processing frees human analysts to concentrate on higher-level interpretation and decision-making (Mitchell, 2019). This hybrid approach harnesses AI's efficiency and scalability while ensuring that human oversight remains in place for crucial judgment calls, thereby maintaining a balance between automated processes and human expertise.

This reimagined intelligence cycle provides a broad framework that can be flexibly utilised to assign tasks appropriately across its three streams. Importantly, the practical effectiveness of the augmented intelligence cycle hinges on responsible and precise task division. Hence, only those tasks suitable for automation should be automated (Davenport & Kirby, 2016; Hagendorff & Wezel, 2020). This necessitates a careful approach to HMCD, ensuring analysts are adequately trained and knowledgeable about AI systems' capabilities, expectations, and appropriate interaction methods. Concurrently, the system's design should continuously evolve to support human users efficiently and robustly.

Chapter 6

Conclusion

This dissertation set out to explore the novel intersection of AI and the intelligence cycle, highlighting its significant potential while addressing the vulnerabilities introduced by AAI. As the modern intelligence landscape rapidly evolves, AI stands at the forefront of this transformation, promising to revolutionise data collection, processing, and analysis. However, with great power comes great responsibility—and risk. Through the examination of several case studies, focusing on both the potential and risks of AI in automating steps of the intelligence cycle, it becomes apparent that the traditional intelligence cycle inadequately accommodates this new hybrid form. As such, the aim of this dissertation was to propose a revised model that integrates and compartmentalises human oversight and advanced AI systems to augment efficiency and accuracy while mitigating the risks of manipulation.

To address this aim comprehensively, the dissertation examined three pivotal stages of the intelligence cycle: collection, processing, and analysis. These stages were selected based on their high potential for automation and their critical impact on the overall intelligence process. Each case explores a specific application of AI and a selection of adversarial vulnerabilities, ensuring a holistic view of the challenges of automating parts of the intelligence cycle.

The first case examined the integration of AI in data collection through autonomous UAVs equipped with CNNs for GEOINT. These UAVs revolutionise data collection by automatically identifying, tracking, and categorising objects, ensuring that only the most relevant data is gathered. Yet, the case highlighted that this automation is vulnerable to evasion attacks, such as adversarial patches, where adversaries manipulate the input data to fool the system. This vulnerability means that, despite their impressive capabilities, UAVs can have their data integrity compromised, which is fundamental for the accuracy of subsequent steps in the intelligence cycle.

The second case delved into the use of NLP for automated data preprocessing. NLP significantly boosts the efficiency and accuracy of tasks like translation, error detection, and misinformation identification. However, the automation in this stage is again not without its risks. The case revealed susceptibility to poisoning attacks, where adversaries corrupt the training data to distort the system's outputs. This manipulation can severely affect the integrity of the processed data, which is essential for the accuracy of the subsequent intelligence analysis.

Finally, the third case investigated the integration of GenAI systems in the analysis phase, where AI assists in generating potential COAs and providing strategic insights. GenAI systems can mitigate cognitive biases and offer diverse perspectives, enhancing hypothesis generation. Yet, this phase also faces significant challenges. The case highlighted the threat of DDoD attacks, which target both the AI system and human analysts to deceive the decision-making process. This dual vulnerability can undermine the reliability of the analysis, the critical foundation on which informed strategic decisions are made.

An analysis of each case reveals the significant impact AI can have on the intelligence domain, underscoring the need for a balanced approach to its integration within the intelligence cycle. Indeed, AI's vulnerabilities can undermine reliability and integrity at multiple stages, jeopardising decisions that rely on accurate intelligence outputs. While AI offers significant benefits in terms of efficiency and capability, its integration must be carefully managed. Addressing the challenges of this new hybrid form requires a nuanced approach, determining tasks suitable for automation on an individual basis and considering the specific context and requirements. This would ensure that AI is employed where it can genuinely enhance operations while mitigating undue risks.

Yet, to optimally accommodate such a robust collaboration between humans and AI-enabled entities, the traditional cycle is seemingly insufficient. Humans and computers possess different qualities and communicate in inherently different languages, fundamentally changing the way in which intelligence is practised. AI is no longer just a technology for performing tedious tasks to save time and resources; it now aims to augment and improve human capabilities. This shift creates hybrid teams characterised by tasks performed by humans, AI, and the synthesis between them. When executed correctly and integrated optimally, this collaboration can generate outputs beyond the scope attainable by humans alone. Therefore, if we wish for AI to run through the same cycle in a distinguished yet integrated manner, a new framework is needed to accommodate this integration.

To effectively address the complexities and changed nature of a hybrid form that includes human agents, I have proposed an augmented intelligence cycle as a suitable new framework. By clearly delineating the roles and responsibilities of human and AI agents, and ensuring continuous synthesis and collaboration, my framework—consisting of two concentric and simultaneously running cycles—aims to leverage the strengths of both while mitigating their respective weaknesses. This balanced approach not only enhances the efficiency and effectiveness of intelligence operations but also provides room for continuous resilience against the evolving threats posed by AAI.

The adoption of the cycle aims to encourage the development of strong standard operating procedures (SOPs) that clearly outline which tasks should be automated and which should not, the level of autonomy given to each party, the associated responsibilities, and what human agents can realistically expect from their new technical counterpart. These procedures are not aimed at repetition or making human agents operate on autopilot; rather they serve to instil a consistent practice of vigilance against potential vulnerabilities and effective HMCD. The augmented intelligence cycle facilitates this by establishing a standardised, conscientious approach to collaborating with a new type of colleague in the intelligence domain. This integration ensures that both human and AI agents work together effectively, accommodating the optimal integration of AI in the iterative intelligence cycle as a whole.

For many, the concept of an automated intelligence cycle might seem far-fetched. However, the field is advancing at an unprecedented pace. To illustrate, when this dissertation was being conceptualised, AI-powered warfare was still confined to theoretical exploration within academic literature. Now, the Israel-Palestine conflict is described as an "AI-assisted genocide" (Al Jazeera, 2024). This starkly illustrates the dramatic and potentially perilous integration of AI into modern conflicts and intelligence efforts, with theory now quickly becoming reality.

As AI technology advances and the volume of available data grows, the question of integrating AI into high-stakes environments has shifted from 'if' to 'when'. AI's capacity to swiftly and accurately process vast amounts of data provides intelligence agencies with unprecedented strategic advantages, enabling them to make informed decisions with remarkable speed. This capability is especially crucial in modern warfare, where data overload and the rapid pace of decision-making can be decisive factors. However, to harness AI's full potential and turn it into a formidable asset rather than a liability, the intelligence cycle and the people within it must evolve along with it. This evolution requires embracing innovation, adapting to new methodologies, and fostering a culture of continuous learning and vigilance. As for now, the near future of intelligence hinges on our human ability to seamlessly integrate our own expertise with AI's capabilities, creating a synergy that is greater than the sum of its parts.

By adopting and continuously reimagining an intelligence cycle that is resilient, adaptive, and forward-thinking, we ensure that AI becomes a powerful ally in safeguarding our security and advancing our strategic interests – paving the way for a world in which AI fosters security rather than undermines it.

6.1 Limitations & Future research

This dissertation addresses a critical research gap by being among the first to explore the implications of AI and AAI for the partial automation of the intelligence cycle. While existing literature is abundant on AI (e.g., Goodfellow et al., 2016; Huang, 2023; Russell & Norvig, 2019) and AAI (e.g., Goodfellow et al., 2018; Hossain et al., 2024; Papernot et al., 2016), as well as the intelligence cycle (e.g., Clark, 2022; Kent, 1966; Lowenthal, 2020), these areas had yet to intersect. While some articles discuss the influence of AI on individual elements of the intelligence cycle (e.g., Erskine & Miller, 2024; Logan, 2024; Rashid et al., 2023), they do not address the impact on the workability of the intelligence cycle as a whole, nor do they consider the significant threat posed by AAI. However, in addressing this research gap, this dissertation also encounters several limitations.

One limitation is the nascent state of AI automation within intelligence processes. AI technology, especially AAI, is rapidly evolving, and its integration into intelligence seems to still be in its infancy. Much of the current understanding is based on theoretical models and simulations rather than real-world applications, highlighting the need for ongoing monitoring and adjustment, as these technologies are deployed.

The secretive nature of intelligence work also poses significant research obstacles. Access to primary data on AI applications in intelligence is highly restricted due to national security concerns. Consequently, this study relies heavily on secondary data from academic literature, which cannot fully substitute for the nuanced insights that primary data could provide. Future research should therefore focus on empirical data outside simulated environments to capture real-world complexities.

The diverse nature of adversarial attacks also poses challenges. While this research primarily focused on several attack categories such as evasion attacks or data poisoning attacks, AAI encompasses a wide range of attacks, each with unique implications and countermeasures – and these types of attacks are developing quickly. Hence, with each new defence, a new way to circumvent it also rapidly surfaces. Future studies should therefore broaden the scope to include other forms of adversarial attacks and investigate their impact on different AI models and applications within the intelligence cycle.

While this dissertation focuses on setting out the risks associated with AAI in the intelligence cycle, it is imperative that future research builds upon these findings to develop robust defence mechanisms. Future research should aim to translate these identified risks into actionable defence strategies, focusing on creating resilient AI systems. This involves not only the development of technical solutions but also the establishment of comprehensive training programs for intelligence personnel. By equipping them with the necessary skills and knowledge to recognise and mitigate these threats, the overall security and reliability of AI-integrated intelligence operations can be significantly enhanced. To achieve this, interdisciplinary collaboration will be crucial, combining insights from fields such as cybersecurity, ML, ethics, and psychology to address the multifaceted nature of an automated intelligence cycle.

In conclusion, while AI holds significant promise for revolutionising intelligence operations, its integration presents challenges requiring careful consideration and ongoing research. Addressing the limitations of current AI applications and focusing on holistic, ethical, and adaptive solutions will harness AI's potential while mitigating risks, ultimately enhancing the effectiveness and security of intelligence operations.

References[2]

Abraham, Y. (2024, April 3). 'Lavender': The AI machine directing Israel's bombing spree in Gaza. *+972 Magazine*. https://www.972mag.com/lavender-ai-israeli-army-gaza/

Abrishami, M., Dadkhah, S., Pinto Neto, E. C., Xiong, P., Iqbal, S., Ray, S., & Ghorbani, A. A. (2022). Classification and analysis of adversarial machine learning attacks in IoT: A label flipping attack case study. *In 2022 32nd Conference of Open Innovations Association (FRUCT)* (pp. 3-14). IEEE. https://doi.org/10.23919/FRUCT56874.2022.9953823

Aggarwal, C. C. (2018). *Neural networks and deep learning: A Textbook*. Springer.

Aghakhani, H., Schönherr, L., Eisenhofer, T., Kolossa, D., Holz, T., Kruegel, C., & Vigna, G. (2023). Venomave: Targeted poisoning against speech recognition. *In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 404-417). Raleigh, NC, USA. https://doi.org/10.1109/SaTML54575.2023.00035

Airbus. (n.d.). *Fortion Massive Intelligence*. Retrieved May 20, 2024, from https://intelligence.airbus.com/industries/defence/joint-isr/fortion-massive-intelligence/

Al Jazeera. (2024, April 4). 'AI-assisted genocide': Israel reportedly used database for Gaza kill lists. *Al Jazeera*. https://www.aljazeera.com/news/2024/4/4/ai-assisted-genocide-israel-reportedly-used-database-for-gaza-kill-lists

Alfayoumi, M., Alshraideh, M., Leiner, M., & AlDajani, I. (2021). Machine Learning predictions for the advancement of the online education in the higher education institutions in

[2] Disclaimer on the use of generative AI: In the creation of this dissertation, generative AI was employed strictly for inspiration, translation, and minor proofreading tasks. Every idea, piece of writing, structural and design choices, or any other fundamental aspect of this dissertation is a product of my own independent effort and creativity.

Jordan. *Journal of Hunan University, 48*(9). http://jonuns.com/index.php/journal/article/download/727/724

Ali, A., Azevedo-Sa, H., Tilbury, D. M., & Robert, L. P. (2022). Heterogeneous human–robot task allocation based on artificial trust. *Scientific Reports*, *12*(1). https://doi.org/10.1038/s41598-022-19140-5

Alpaydin, E. (2020). *Introduction to Machine Learning* (4th ed.). MIT Press.

Alzetta, F., Giorgini, P., Najjar, A., Schumacher, M. I., & Calvaresi, D. (2020). In-time explainability in multi-agent systems: Challenges, opportunities, and roadmap. In D. Calvaresi, A. Najjar, M. Winikoff, & K. Främling (Eds.), *Explainable, transparent autonomous agents and multi-agent systems (EXTRAAMAS 2020). Lecture notes in computer science* (Vol. 12175). Springer, Cham. https://doi.org/10.1007/978-3-030-51924-7_3

Amaratunga, T. (2023). *Understanding large language models: Learning Their Underlying Concepts and Technologies*. Apress.

Amoroso, D., & Tamburrini, G. (2020). Autonomous weapons systems and meaningful human control: ethical and legal issues. *Current Robotics Reports*, *1*(4), 187–194. https://doi.org/10.1007/s43154-020-00024-3

Andrews, R. W., Lilly, J. M., Srivastava, D., & Feigh, K. M. (2022). The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science, 24*(2), 129–175. https://doi.org/10.1080/1463922x.2022.2061080

Ayub, M. A., Johnson, W. A., Talbert, D. A., & Siraj, A. (2020). Model evasion attack on intrusion detection systems using adversarial machine learning. *In 2020 54th Annual Conference on Information Sciences and Systems (CISS)* (pp. 1-6). IEEE. https://doi.org/10.1109/CISS48834.2020.1570617116

Baber, C., McCormick, E., & Apperley, I. (2021). Understanding The Problem Of Explanation When Using Artificial Intelligence In Intelligence Analysis. *In Centre for Research and Evidence on Security Threats*. https://crestresearch.ac.uk/resources/understanding-the-problem-of-explanation-when-using-ai-in-intelligence-analysis/

Baniecki, H., & Biecek, P. (2024). Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion*, 102303. https://doi.org/10.1016/j.inffus.2024.102303

Bauer, M. W., & Schiele, B. (2024). *AI and Common Sense: Ambitions and Frictions*. Taylor & Francis.

Belton, I., & Dhami, M. K. (2019). Cognitive biases and debiasing in intelligence analysis. *In M. J. Turner (Ed.), Advanced research in intelligence and national security* (pp. 548-560). Routledge. https://doi.org/10.4324/9781315658353-43

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Billal, B., Fonseca, A., & Sadat, F. (2016). Efficient natural language pre-processing for analyzing large data sets. *In 2016 IEEE International Conference on Big Data* (pp. 3864-3871). Washington, DC, USA: IEEE. https://doi.org/10.1109/BigData.2016.7841060

Blanchard, A., & Taddeo, M. (2023). The Ethics of Artificial Intelligence for Intelligence Analysis: a Review of the Key Challenges with Recommendations. *Digital Society, 2*(1). https://doi.org/10.1007/s44206-023-00036-4

Breakspear, A. (2013). A New Definition of Intelligence. *Intelligence & National Security, 28*(5), 678–693. https://doi.org/10.1080/02684527.2012.699285

Brink, N., Kamphuis, Y., Maas, Y., Jansen-Ferdinandus, G., Van Stijn, J., Poppink, B., De Haan, P., & Chiscop, I. (2023). Adversarial AI in the cyber domain. *TNO*. https://publications.tno.nl/publication/34640702/FYQbTu/TNO-2023-R10292-EN.pdf

Buchanan, J., & Kock, N. (2001). Information overload: A decision making perspective. In M. Köksalan & S. Zionts (Eds.), *Multiple criteria decision making in the new millennium* (Vol. 507, pp. 49-59). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-56680-6_4

Burda, P., Allodi, L., & Zannone, N. (2024). Cognition in Social Engineering Empirical Research: A Systematic Literature review. *ACM Transactions on Computer-human Interaction, 31*(2), 1–55. https://doi.org/10.1145/3635149

Burke, P. (2021). Intelligence and national Security: The National Security Problematique. *In M. Clarke, A. Henschke, M. Sussex, & T. Legrand (Eds.), The Palgrave Handbook of National Security* (pp. 351–370). Palgrave Macmillan. https://doi.org/10.1007/978-3-030-53494-3_15

Calegari, R., Ciatto, G., Mascardi, V., & Omicini, A. (2021). Logic-based technologies for multi-agent systems: A systematic literature review. *Autonomous Agents and Multi-Agent Systems, 35*(1), 1. https://doi.org/10.1007/s10458-020-09478-3

Calvaresi, D., Cid, Y. D., Marinoni, M., Dragoni, A. F., Najjar, A., & Schumacher, M. (2021). Real-time multi-agent systems: rationality, formal model, and empirical results. *Autonomous Agents and Multi-Agent Systems, 35*(1). https://doi.org/10.1007/s10458-020-09492-5

Carlini, N., Jagielski, M., Choquette-Choo, C. A., Paleka, D., Pearce, W., Anderson, H., Terzis, A., Thomas, K., & Tramèr, F. (2023). Poisoning web-scale training datasets is practical. *arXiv*. https://doi.org/10.48550/arXiv.2302.10149

Chen, J., Zhang, Y., Liu, C., Chen, K., Zou, Z., & Shi, Z. (2024). Digital-to-physical visual consistency optimization for adversarial patch generation in remote sensing scenes. *IEEE Transactions on Geoscience and Remote Sensing, 62*, 1-17. https://doi.org/10.1109/TGRS.2024.3397678

Chen, Z., Lin, P., Jiang, Z. L., Wei, Z., Yuan, S., & Fang, J. (2021). An illumination modulation-based adversarial attack against automated face recognition system. *In Y. Wu & M. Yung (Eds.), Information security and cryptology. Inscrypt 2020. Lecture notes in computer science* (Vol. 12612). Springer. https://doi.org/10.1007/978-3-030-71852-7_4

Chow, A. R., & Perrigo, B. (2023, February 17). The AI arms race is changing everything. *TIME*. https://time.com/6255952/ai-impact-chatgpt-microsoft-google/

Christa, S., Suma, V., & Madhuri, L. (2013). Data preprocessing using intelligent agents. *In S. Satapathy, S. Udgata, & B. Biswal (Eds.), Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) (Advances in Intelligent Systems and Computing, vol. 199*, pp. 23-31). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35314-7_23

Christie, E. H., Ertan, A., Adomaitis, L., & Klaus, M. (2023). Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability. *AI And Ethics*. https://doi.org/10.1007/s43681-023-00261-0

Chun-Wei Chiang, Zhuoran Lu, Zhuoyan Li, and Ming Yin. (2024). Enhancing AI-Assisted Group Decision Making through LLM-Powered Devil's Advocate. *In 29th International Conference on Intelligent User Interfaces (IUI '24)*, March 18–21, 2024, Greenville, SC, USA. ACM, New York, NY, USA, 17 pages. https://doi.org/10.1145/3640543.3645199

Clark, R. (2022). *Intelligence Analysis: A Target-Centric Approach* (7th ed.). CQ Press.

Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering, 27*(1), 113–118. https://doi.org/10.1017/s1351324920000601

Darnell, J. A., & Gopalkrishnan, S. (2023). Digital information overload: How leaders can strategically use AI to prevent innovation paralysis. In N. Pfeffermann & M. Schaller (Eds.), *New leadership communication—inspire your horizon* (pp. 181–190). Springer, Cham. https://doi.org/10.1007/978-3-031-34314-8_14

Davenport, T. H., & Kirby, J. (2016). *Only humans need apply: Winners & Losers in the Age of Smart Machines.* HarperCollins.

De Visser, E. J., Peeters, M. M. M., Jung, M. F., et al. (2020). Towards a theory of longitudinal trust calibration in human–robot teams. *International Journal of Social Robotics, 12*(2), 459–478. https://doi.org/10.1007/s12369-019-00596-x

Den Hollander, R., Adhikari, A., Tolios, I., van Bekkum, M., Bal, A., Hendriks, S., Kruithof, M., Gross, D., Jansen, N., Perez, G., Buurman, K., & Raaijmakers, S. (2020, September). Adversarial patch camouflage against aerial detection. *In Artificial Intelligence and Machine Learning in Defense Applications II* (Vol. 11543, 115430F). SPIE. https://doi.org/10.1117/12.2575907

Dogo, E. M., Afolabi, O. J., Nwulu, N. I., Twala, B., & Aigbavboa, C. O. (2018). A comparative analysis of gradient descent-based optimization algorithms on convolutional neural networks. In *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)* (pp. 92-99). IEEE. https://doi.org/10.1109/CTEMS.2018.8769211

Du, W., Yuan, T., Zhao, H., & Liu, G. (2024). NWS: Natural textual backdoor attacks via word substitution. *In ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 4680-4684). IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447968

Ehsan, U., Wintersberger, P., Liao, Q. V., Watkins, E. A., Manger, C., Daumé, H., III, Riener, A., & Riedl, M. O. (2022). Human-Centered Explainable AI (HCXAI): beyond opening the Black-Box of AI. *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. https://doi.org/10.1145/3491101.3503727

Emilien, A., Thomas, C., & Thomas, H. (2021). UAV & satellite synergies for optical remote sensing applications: A literature review. *Science of Remote Sensing, 3*, 100019. https://doi.org/10.1016/j.srs.2021.100019

Erskine, T., & Miller, S. E. (2024). AI and the decision to go to war: future risks and opportunities. *Australian Journal of International Affairs*, *78*(2), 135–147. https://doi.org/10.1080/10357718.2024.2349598

Eschenbach, V., & Warren, J. R. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology, 34*(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

Evangelista, J. R. G., Sassi, R. J., Romero, M., & Napolitano, D. M. R. (2020). Systematic Literature Review to Investigate the Application of Open Source Intelligence (OSINT) with Artificial Intelligence. *Journal of Applied Security Research, 16*(3), 345–369. https://doi.org/10.1080/19361610.2020.1761737

Evans, G. (2009). Rethinking military intelligence failure – putting the wheels back on the intelligence cycle. *Defence Studies, 9*(1), 22–46. https://doi.org/10.1080/14702430701811987

Fabri, L., Häckel, B., Oberländer, A. M., Rieg, M., & Stohr, A. (2023). Disentangling Human-AI Hybrids. *Business & Information Systems Engineering, 65*(6), 623–641. https://doi.org/10.1007/s12599-023-00810-1

Gan, L., Li, J., Zhang, T., Li, X., Meng, Y., Wu, F., Yang, Y., Guo, S., & Fan, C. (2022, July). Triggerless backdoor attack for NLP tasks with clean labels. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2942–2952). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.214

Gandham, S. (2020, April 6). Is it a bird or a plane? One pixel attack for fooling deep neural networks. *Medium*. https://medium.com/machine-intelligence-and-deep-learning-lab/is-it-a-bird-or-a-plane-one-pixel-attack-for-fooling-deep-neural-networks-ac9459efc781

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences, 29*, 17–23. https://doi.org/10.1016/j.cobeha.2018.12.010

George, A. L., & Bennett, A. (2005). *Case studies and theory development in the social sciences*. The MIT Press.

Gerring, J. (2016). *Case study research: Principles and Practices* (2nd ed.). Cambridge University Press.

Gill, P., & Phythian, M. (2013). From intelligence cycle to web of intelligence. *In M. Phythian, Understanding the Intelligence Cycle* (pp. 21–42). Routledge.

Gnanasambandam, A., Sherman, A. M., & Chan, S. H. (2021). Optical adversarial attack. *In 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)* (pp. 92-101). IEEE. https://doi.org/10.1109/ICCVW54120.2021.00016

Goldman Sachs. (2023, August 1). AI investment forecast to approach $200 billion globally by 2025. *Goldman Sachs*. https://www.goldmansachs.com/intelligence/pages/ai-investment-forecast-to-approach-200-billion-globally-by-2025.html

Goldwasser, S., Kim, M. P., Vaikuntanathan, V., & Zamir, O. (2022). Planting undetectable backdoors in machine learning models. *arXiv*. https://doi.org/10.48550/arxiv.2204.06974

Goodfellow, I. J., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.

Goodfellow, I. J., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM, 61*(7), 56–66. https://doi.org/10.1145/3134599

Grigorieva, N. M., & Petrenko, S. A. (2023). Known adversarial attacks by 2023. In *2023 Seminar on Information Systems Theory and Practice (ISTP)* (pp. 32-36). Saint Petersburg, Russian Federation. https://doi.org/10.1109/ISTP60767.2023.10427453

Gu, J., Wang, Z., Kuen, J., Ma, L., Shahroudy, A., Shuai, B., Liu, T., Wang, X., Wang, G., Cai, J., & Chen, T. (2018). Recent advances in convolutional neural networks. *Pattern Recognition*, *77*, 354–377. https://doi.org/10.1016/j.patcog.2017.10.013

Guest, P. (2024, March 6). New AI video tools increase worries of deepfakes ahead of elections. *Al Jazeera*. https://www.aljazeera.com/economy/2024/3/6/new-ai-video-tools-increase-worries-of-deepfakes-ahead-of-elections

Guo, W., Tondi, B., & Barni, M. (2022). An overview of backdoor attacks against deep neural networks and possible defences. *IEEE Open Journal of Signal Processing, 3*, 261–287. https://doi.org/10.1109/ojsp.2022.3190213

Hagendorff, T., & Wezel, K. (2020). 15 challenges for AI: Or what AI (currently) can't do. *AI & Society, 35*(2), 355–365. https://doi.org/10.1007/s00146-019-00886-y

Hao, J., Von Davier, A. A., Yaneva, V., Lottridge, S., Von Davier, M., & Harris, D. J. (2024). Transforming Assessment: The impacts and implications of large language models and Generative AI. *Educational Measurement*, *43*(2), 16–29. https://doi.org/10.1111/emip.12602

Henrico, S., & Putter, D. (2024). Intelligence Collection Disciplines—A Systematic Review. *Journal of Applied Security Research*, 1–25. https://doi.org/10.1080/19361610.2023.2296765

Heuer, R. J., Jr. (1999). *Psychology of intelligence analysis*. Center for the Study of Intelligence, Central Intelligence Agency. https://archive.org/details/PsychologyOfIntelligenceAnalysis

Hoppa, M. A., Debb, S. M., Hsieh, G., & KCa, B. (2019). TwitterOSINT: Automated Open Source Intelligence Collection, Analysis & Visualization Tool. *Annual Review of Cybertherapy And Telemedicine 2019*, 121-128.

Hossain, M., Afrin, R., & Biswas, M. (2024). A Review on Attacks against Artificial Intelligence (AI) and Their Defence Image Recognition and Generation Machine Learning, Artificial Intelligence. *Control Systems and Optimization Letters, 2*(1), 52-59. https://doi.org/10.59247/csol.v2i1.73

Huang, M., Liu, A., Xiong, N. N., & Wu, J. (2021). A UAV-assisted ubiquitous trust communication system in 5G and beyond networks. *IEEE Journal on Selected Areas in Communications, 39*(11), 3444-3458. https://doi.org/10.1109/JSAC.2021.3088675

Huang, Y. (2023). The future of generative AI: How GenAI would change human-computer co-creation in the next 10 to 15 years. *In Companion Proceedings of the Annual Symposium on Computer-Human Interaction in Play (CHI PLAY Companion '23)* (pp. 1-4). ACM. https://doi.org/10.1145/3573382.3616033

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., & Mądry, A. (2019). Adversarial examples are not bugs, they are features. *Neural Information Processing Systems, 32*, 125–136. https://papers.nips.cc/paper/8307-adversarial-examples-are-not-bugs-they-are-features.pdf

Intel AI. (2019, May 28). The rise in Computing Power: Why ubiquitous artificial intelligence is now a reality. *Forbes*. https://www.forbes.com/sites/intelai/2018/07/17/the-rise-in-computing-power-why-ubiquitous-artificial-intelligence-is-now-a-reality/

Irfan, M. M., Ali, S., Yaqoob, I., & Zafar, N. (2021). Towards deep learning: A review on adversarial attacks. In *2021 International Conference on Artificial Intelligence (ICAI)* (pp. 91-96). IEEE. https://doi.org/10.1109/ICAI52203.2021.9445247

Jensen, C., McElreath, D., & Graves, M. (2022). *Introduction to Intelligence Studies* (3rd ed.). Routledge.

Jiang, M. (2020). Improving situational awareness with collective artificial intelligence over knowledge graphs. In *Artificial Intelligence and Machine Learning for Multi-Domain Operations Applications II*. SPIE. https://doi.org/10.1117/12.2556746

Jiang, W., Li, H., Gong, L., Yang, H., & Lu, R. (2020). Accelerating poisoning attack through Momentum and Adam algorithms. *In 2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)* (pp. 1-6). IEEE. https://doi.org/10.1109/VTC2020-Fall49728.2020.9348449

Jouhari, M., Al-Ali, A., Baccour, E., Mohamed, A., Erbad, A., Guizani, M., & Hamdi, M. (2022). Distributed CNN Inference on Resource-Constrained UAVs for Surveillance Systems: Design and Optimization. *IEEE Internet of Things Journal, 9*(2), 1227–1242. https://doi.org/10.1109/jiot.2021.3079164

Kahn, D. (2001). An historical theory of intelligence. *Intelligence & National Security, 16*(3), 79–92. https://doi.org/10.1080/02684520412331306220

Kahneman, D. (2011). *Thinking, fast and slow*. Penguin UK.

Kaswan, K. S., Dhatterwal, J. S., Malik, K., & Baliyan, A. (2023). Generative AI: A review on models and applications. In *2023 International Conference on Communication, Security and Artificial Intelligence (ICCSAI)* (pp. 699-704). IEEE. https://doi.org/10.1109/ICCSAI59793.2023.10421601.

Kent, S. (1966). *Strategic Intelligence for American World Policy*. Princeton University Press.

Kerr, A. (2023). Researching national security intelligence: Multidisciplinary approaches: Edited by Stephen Coulthart, Michael Landon-Murray, and Damien Van Puyvelde. Washington, DC: Georgetown University Press. *Intelligence and National Security, 38*(5), 838–843. https://doi.org/10.1080/02684527.2023.2170354

King, A. (2024). Digital targeting: artificial intelligence, data, and military intelligence. *Journal of Global Security Studies*, *9*(2). https://doi.org/10.1093/jogss/ogae009

Krizan, L. (1999). *Intelligence essentials for everyone*. Joint Military Intelligence College Press. https://archive.org/details/intelligenceessential00kriz

Labadze, L., Grigolia, M., & Machaidze, L. (2023). Role of AI chatbots in education: systematic literature review. *International Journal Of Educational Technology in Higher Education, 20*(1). https://doi.org/10.1186/s41239-023-00426-1

Legaspi, R., Xu, W., Konishi, T., Wada, S., Kobayashi, N., Naruse, Y., & Ishikawa, Y. (2024). The sense of agency in human-AI interactions. *Knowledge-based Systems*, *286*, 111298. https://doi.org/10.1016/j.knosys.2023.111298

Li, Z., Liu, F., Yang, W., Peng, S., & Zhou, J. (2022). A Survey of Convolutional Neural Networks: Analysis, applications, and Prospects. *IEEE Transactions on Neural Networks and Learning Systems, 33*(12), 6999–7019. https://doi.org/10.1109/tnnls.2021.3084827

Li, Z., Xia, P., Tao, R., Niu, H., & Li, B. (2023). A new perspective on stabilizing GANs training: direct adversarial training. *IEEE Transactions on Emerging Topics in Computational Intelligence*, *7*(1), 178–189. https://doi.org/10.1109/tetci.2022.3193373

Li, G., Zhou, X., & Cao, L. (2021). AI meets database: AI4DB and DB4AI. In *Proceedings of the 2021 International Conference on Management of Data (SIGMOD '21)* (pp. 2859–2866). Association for Computing Machinery. https://doi.org/10.1145/3448016.3457542

Lian, J., Mei, S., Zhang, S., & Ma, M. (2022). Benchmarking adversarial patch against aerial detection. *IEEE Transactions on Geoscience and Remote Sensing, 60*, 1-16. https://doi.org/10.1109/TGRS.2022.3225306

Liben, F., Abera, W., Chernet, M. T., Ebrahim, M., Tilaye, A., Erkossa, T., Degefie, D. T., Mponela, P., Kihara, J., & Tamene, L. (2024). Site-specific fertilizer recommendation using data driven machine learning enhanced wheat productivity and resource use efficiency. *Field Crops Research, 313*, 109413. https://doi.org/10.1016/j.fcr.2024.109413

Logan, S. (2024). Tell me what you don't know: large language models and the pathologies of intelligence analysis. *Australian Journal of International Affairs*, *78*(2), 220–228. https://doi.org/10.1080/10357718.2024.2331733

Lowenthal, M. M. (2020). *Intelligence: From Secrets to Policy* (8th ed.). SAGE Publications.

Liu, X., Song, H., & Liu, A. (2021). Intelligent UAVs trajectory optimization from Space-Time for data collection in social networks. *IEEE Transactions on Network Science and Engineering*, *8*(2), 853–864. https://doi.org/10.1109/tnse.2020.3017556

Lu, Z., Li, Z., Chiang, C. W., & Yin, M. (2023). Strategic adversarial attacks in AI-assisted decision making to reduce human trust and reliance. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (pp. 3020-3028).

MacQuarrie, C. (2010). Consciousness raising. *In A. J. Mills, G. Durepos, & E. Wiebe (Eds.), Encyclopedia of case study research* (pp. 213–215). Sage.

Makridakis, S. (2017). The forthcoming Artificial Intelligence (AI) revolution: Its impact on society and firms. *Futures*, *90*, 46–60. https://doi.org/10.1016/j.futures.2017.03.006

Marzouk, A., Hasso, B. A., Tchappi, I., Al-Kuwatli, B., & Najjar, A. (2023). BlueData: AI Assisted Primary Data Collection System for Conflict Zones. *In Proceedings of the 11th International Conference on Human-Agent Interaction*. Association for Computing Machinery. https://doi.org/10.1145/3623809.3623946

Marulli, F., Verde, L., & Campanile, L. (2021). Exploring data and model poisoning attacks to Deep Learning-Based NLP systems. *Procedia Computer Science, 192*, 3570–3579. https://doi.org/10.1016/j.procs.2021.09.130

Meske, C., & Bunde, E. (2020). Transparency and trust in human-AI interaction: The role of model-agnostic explanations in computer vision-based decision support. In H. Degen & L. Reinerman-Jones (Eds.), *Artificial intelligence in HCI. HCII 2020. Lecture notes in computer science* (Vol. 12217). Springer, Cham. https://doi.org/10.1007/978-3-030-50334-5_4

Mitchell, K., Mariani, J., Routh, A., Keyal, A., & Mirkow, A. (2019). The future of intelligence analysis. *In Deloitte Insights*. https://www.deloitte.com/ke/en/our-thinking/insights/industry/government-public-services/artificial-intelligence-impact-on-future-intelligence-analysis.html

Morales, E. F., & Escalante, H. J. (2022). A brief introduction to supervised, unsupervised, and reinforcement learning. *In A. Torres-García, C. Reyes-García, L. Villaseñor-Pineda, & O. Mendoza-Montoya (Eds.), Biosignal Processing and Classification Using Computational Learning and Intelligence* (pp. 111–129). Elsevier. https://doi.org/10.1016/b978-0-12-820125-1.00017-8

Nam, W., Kim, K., Moon, H., Noh, H., Park, J., & Kil, H. (2024). RISOPA: Rapid imperceptible strong one-pixel attacks in deep neural networks. *Mathematics, 12*(7), 1083. https://doi.org/10.3390/math12071083

Neupane, D., & Seok, J. (2020). A review on Deep Learning-Based Approaches for Automatic Sonar Target Recognition. *Electronics, 9*(11), 1972. https://doi.org/10.3390/electronics9111972

Nguyen-Son, H.Q., Thao, T. P., Hidano, S., Bracamonte, V., Kiyomoto, S., & Yamaguchi, R. S. (2021). OPA2D: One-pixel attack, detection, and defense in deep neural networks. *2021 International Joint Conference on Neural Networks (IJCNN)*, Shenzhen, China, 1-10. https://doi.org/10.1109/IJCNN52387.2021.9534332

Okamura, K., & Yamada, S. (2020). Empirical Evaluations of Framework for Adaptive trust Calibration in Human-AI Cooperation. *IEEE Access, 8*, 220335–220351. https://doi.org/10.1109/access.2020.3042556

Omand, D. (2023). The historical backdrop. *In S. Stenslie, L. Haugom, & B. Vaage, Intelligence Analysis in the Digital Age*. Routledge.

OpenAI. (n.d.). Sora. *OpenAI*. Retrieved June 11, 2024, from https://openai.com/index/sora/

Palantir. (n.d.-a). *Gotham*. Retrieved May 1, 2024, from https://www.palantir.com/platforms/gotham/

Palantir. (n.d.-b). *Palantir Defense Solutions*. Retrieved December 30, 2023, from https://www.palantir.com/offerings/defense/

Palantir. (2020). AI-Enabled Operations. *In Palantir Technologies*. https://www.palantir.com/assets/xrfr7uokpv1b/3A0y10xksgXENvRMNaAsUu/ed8f7f1ed534c0101f64536a85f7297b/Gotham_AI-Enabled_Operations_White_Paper.pdf

Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)* (pp. 372-387). IEEE. https://doi.org/10.1109/EuroSP.2016.36

Pastor-Galindo, J., Nespoli, P., Mármol, F. G., & Pérez, G. M. (2020). The not yet exploited goldmine of OSINT: Opportunities, open challenges and future trends. *IEEE Access, 8*, 10282–10304. https://doi.org/10.1109/access.2020.2965257

Pherson, R. H., & Heuer, R. J., Jr. (2019). *Structured analytic techniques for intelligence analysis*. CQ Press.

Phythian, M. (2013). Introduction: Beyond the Intelligence Cycle? *In Understanding the Intelligence Cycle* (pp. 1–9). Routledge.

Qi, F., Chen, Y., Zhang, X., Li, M., Liu, Z., & Sun, M. (2021, November). Mind the style of text! Adversarial and backdoor attacks based on text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 4569–4580). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.374

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. '., . . . Wellman, M. (2019). Machine behaviour. *Nature*, *568*(7753), 477–486. https://doi.org/10.1038/s41586-019-1138-y

Rai, A. (2019). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rao, Y., Yi, Y., Nartey, O. T., & Jan, S. U. (2023). Relevance gradient descent for parameter optimization of image enhancement. *Computers & Graphics, 117*, 124-133. https://doi.org/10.1016/j.cag.2023.10.016

Rashid, A. B., Kausik, A. K., Sunny, A. a. H., & Bappy, M. H. (2023). Artificial intelligence in the military: An overview of the capabilities, applications, and challenges. *International Journal of Intelligent Systems*, *2023*, 1–31. https://doi.org/10.1155/2023/8676366

Räuker, T., Ho, A., Casper, S., & Hadfield-Menell, D. (2023). Toward transparent AI: A survey on interpreting the inner structures of deep neural networks. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (pp. 464-483). IEEE. https://doi.org/10.1109/SaTML54575.2023.00039

Regens, J. L. (2019). Augmenting human cognition to enhance strategic, operational, and tactical intelligence. *Intelligence & National Security, 34*(5), 673–687. https://doi.org/10.1080/02684527.2019.1579410

Ren, M., Chen, N., & Qiu, H. (2023). Human-machine Collaborative Decision-making: An Evolutionary Roadmap Based on Cognitive Intelligence. *International Journal Of Social Robotics, 15*(7), 1101–1114. https://doi.org/10.1007/s12369-023-01020-1

Ren, K., Zheng, T., Qin, Z., & Liu, X. (2020). Adversarial attacks and defenses in deep learning. *Engineering*, *6*(3), 346–360. https://doi.org/10.1016/j.eng.2019.12.012

Rizk, Y., Awad, M., & Tunstel, E. W. (2018). Decision making in multiagent systems: A survey. *IEEE Transactions on Cognitive and Developmental Systems, 10*(3), 514-529. https://doi.org/10.1109/TCDS.2018.2840971

Rohozinski, R. (2023, May 25). *AI is poised to revolutionize intelligence and risk analysis*. Centre for International Governance Innovation. https://www.cigionline.org/articles/ai-is-poised-to-revolutionize-intelligence-and-risk-analysis/

Russell, S., & Norvig, P. (2019). *Artificial intelligence: A Modern Approach*. Pearson Higher Education.

Sadiku, M.N.O., & Musa, S.M. (2021). Augmented Intelligence. *In A Primer on Multiple Intelligences* (pp. 191–199). Springer. https://doi.org/10.1007/978-3-030-77584-1_15

Sarcia, S. A., & Colo, G. (2023). Organizing Structures and Information for Developing AI-enabled Military Decision-Making Systems. https://doi.org/10.1109/techdefense59795.2023.10380925

Sayles, A., Hooda, A., Gupta, M., Chatterjee, R., & Fernandes, E. (2021). Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. *In 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 14661-14670). IEEE. https://doi.org/10.1109/CVPR46437.2021.01443

Schwartz, O. (2019, November 25). In 2016, Microsoft's racist chatbot revealed the dangers of online conversation. *IEEE Spectrum*. https://spectrum.ieee.org/in-2016-microsofts-racist-chatbot-revealed-the-dangers-of-online-conversation

Schwartz, P. J., O'Neill, D., Bentz, M. E., Brown, A., Doyle, B., Liepa, O. C., Lawrence, R., & Hull, R. D. (2020). AI-enabled wargaming in the military decision making process. *In Artificial Intelligence and Machine Learning for Multi-Domain Operations* (Vol. 11413). SPIE. https://doi.org/10.1117/12.2560494

Seawright, J., & Gerring, J. (2008). Case selection techniques in case study research. *Political Research Quarterly, 61*(2), 294–308. https://doi.org/10.1177/1065912907313077

Stake, R. E. (2005). Qualitative Case Studies. *In N. K. Denzin & Y. S. Lincoln (Eds.), The Sage Handbook of Qualitative Research* (3rd ed., pp. 443–466). Sage Publications.

Steen, M., Van Diggelen, J., Timan, T., & Van Der Stap, N. (2022). Meaningful human control of drones: exploring human–machine teaming, informed by four different ethical perspectives. *AI And Ethics, 3*(1), 281–293. https://doi.org/10.1007/s43681-022-00168-2

Stenslie, S., Haugom, L., & Vaage, B. (2023). Introduction: An old activity in a new age. *In Intelligence Analysis in the Digital Age*. Routledge.

Symon, P. B., & Tarapore, A. (2015). Defense intelligence analysis in the age of big data. *Joint Force Quarterly, 79*(4), 04-11.

Tag, B., Van Berkel, N., Verma, S., Zhao, B. Z. H., Berkovsky, S., Kâafar, M. A., Kostakos, V., & Ohrimenko, O. (2023). DDOD: Dual Denial of Decision Attacks on Human-AI Teams. *IEEE Pervasive Computing, 22*(1), 77–84. https://doi.org/10.1109/mprv.2022.3218773

Tian, Z., Cui, L., Liang, J., & Yu, S. (2022). A comprehensive survey on poisoning attacks and countermeasures in machine learning. *ACM Computing Surveys, 55*(8), Article 166, 35 pages. https://doi.org/10.1145/3551636

TNO. (n.d.). Federated learning: privacy-preserving data analysis. Retrieved May 25, 2024, from https://www.tno.nl/en/technology-science/technologies/federated-learning/

Torraco, R. J. (2005). Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review, 4*(3), 356-367. https://doi.org/10.1177/1534484305278283

Torraco, R. J. (2016). Writing integrative literature reviews: Using the past and present to explore the future. *Human Resource Development Review, 15*(4), 404-428. https://doi.org/10.1177/1534484316671606

Truong, L., Jones, C., Hutchinson, B., August, A., Praggastis, B., Jasper, R., Nichols, N., & Tuor, A. (2020). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* (pp. 788–789).

Utomo, A. M., Wijayanto, G. N., Yusfan, M. A., Wardani, P., Poniman, A., Supriyadi, A. A., ... & Arief, S. (2021, October). Geospatial intelligence analysis to support national defense interests. *In 2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)* (pp. 1-8). IEEE. https://doi.org/10.1109/ICACSIS53237.2021.9631348

Van Der Veer, S. N., Riste, L., Cheraghi-Sohi, S., Phipps, D. L., Tully, M. P., Bozentko, K., Atwood, S., Hubbard, A., Wiper, C., Oswald, M., & Peek, N. (2021). Trading off accuracy and explainability in AI decision-making: findings from 2 citizens' juries. *Journal of the American Medical Informatics Association, 28*(10), 2128–2138. https://doi.org/10.1093/jamia/ocab127

Vassilev, A., Oprea, A., Fordyce, A., & Anderson, H. (2024). Adversarial machine learning: A Taxonomy and Terminology of Attacks and Mitigations. *NIST.* https://doi.org/10.6028/nist.ai.100-2e2023

Vössing, M., Kühl, N., Lind, M., & Satzger, G. (2022). Designing transparency for effective Human-AI collaboration. *Information Systems Frontiers, 24*(3), 877–895. https://doi.org/10.1007/s10796-022-10284-3

Wang, D., Jiang, T., Sun, J., Zhou, W., Gong, Z., Zhang, X., Yao, W., & Chen, X. (2022). FCA: Learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack. *Proceedings of the AAAI Conference on Artificial Intelligence, 36*(2), 2414-2422. https://doi.org/10.1609/aaai.v36i2.20141

Wang, J., Xu, C., Guzmán, F., El-Kishky, A., Tang, Y., Rubinstein, B. I., & Cohn, T. (2021). Putting words into the system's mouth: A targeted attack on neural machine translation using monolingual data poisoning. *arXiv preprint arXiv:2107.05243*.

Wang, Z., Ma, J., Wang, X., Hu, J., Qin, Z., & Ren, K. (2022). Threats to training: A survey of poisoning attacks and defenses on machine learning systems. *ACM Computing Surveys, 55*(7), Article 134, 36 pages. https://doi.org/10.1145/3538707

Wei, X., Huang, Y., Sun, Y., & Yu, J. (2024). Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(4), 2348-2363. https://doi.org/10.1109/TPAMI.2023.3330769

Wen, H., & Song, J. (2019). Using big data to enhance the capability of the situational awareness of the battlefield. In Y. Jia, J. Du, & W. Zhang (Eds.), *Proceedings of 2018 Chinese Intelligent Systems Conference*. Lecture Notes in Electrical Engineering, vol 528. Springer, Singapore. https://doi.org/10.1007/978-981-13-2288-4_31

Williamson, T. (2016). Knowing by imagining. *In A. Kind & P. Kung (Eds.), Knowledge through imagination* (online ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198716808.003.0005

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q., & Tang, Y. (2023). A brief overview of ChatGPT: the history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122–1136. https://doi.org/10.1109/jas.2023.123618

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T., & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems, 135*, 364–381. https://doi.org/10.1016/j.future.2022.05.014

Xu, Q., Yang, Z., Zhao, Y., Cao, X., & Huang, Q. (2023). Rethinking label flipping attack: From sample masking to sample thresholding. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 45*(6), 7668-7685. https://doi.org/10.1109/TPAMI.2022.3220849

Xu, Y., Liu, X., Cao, X., Huang, C., Liu, E., Qian, S., Liu, X., Wu, Y., Dong, F., Qiu, C., Qiu, J., Hua, K., Su, W., Wu, J., Xu, H., Han, Y., Fu, C., Yin, Z., Liu, M., ... Zhang, J. (2021). Artificial intelligence: A powerful paradigm for scientific research. *The Innovation, 2*(4), 100179. https://doi.org/10.1016/j.xinn.2021.100179

Yang, S., Li, Q., Wang, P., Lian, Z., Hou, J., & Rao, Y. (2023). Be careful with writing perturbations: This could be a trigger of textual backdoor attacks. *In 2023 IEEE Smart*

World Congress (SWC) (pp. 1-8). IEEE. https://doi.org/10.1109/SWC57546.2023.10449255

Yang, X., Wei, F., Zhang, H., & Zhu, J. (2020). Design and interpretation of universal adversarial patches in face detection. *In A. Vedaldi, H. Bischof, T. Brox, & J. M. Frahm (Eds.), Computer Vision – ECCV 2020. ECCV 2020. Lecture Notes in Computer Science* (Vol. 12362). Springer, Cham. https://doi.org/10.1007/978-3-030-58520-4_11

Yin, S., Wang, L., Shafiq, M., Teng, L., Laghari, A. A., & Khan, M. F. (2023). G2Grad-CAMRL: An object detection and interpretation model based on gradient-weighted class activation mapping and reinforcement learning in remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 16*, 3583-3598. https://doi.org/10.1109/JSTARS.2023.3241405

Yu, S., & Cui, L. (2023). *Security and privacy in federated learning*. Springer Nature.

Yuan, L., & Hu, H. (2023). Model poisoning attack in federated learning via adversarial examples. In *2023 International Seminar on Computer Science and Engineering Technology (SCSET)* (pp. 52-55). New York, NY, USA. https://doi.org/10.1109/SCSET58950.2023.00021

Yuviler, T., & Drachsler-Cohen, D. (2023). One pixel adversarial attacks via sketched programs. *Proceedings of the ACM on Programming Languages, 7*(PLDI), Article 187, 25 pages. https://doi.org/10.1145/3591301

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). "An ideal human": Expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction, 4*(CSCW3), Article 246, 1-25. https://doi.org/10.1145/3432945

Zhou, Y., Rao, B., & Wang, W. (2020). UAV swarm intelligence: Recent advances and future trends. *IEEE Access, 8*, 183856-183878. https://doi.org/10.1109/ACCESS.2020.3028865

Appendix A

	Case 1	*Case 2*	*Case 3*
Keywords	Artificial Intelligence or AI; Adversarial AI; Evasion Attack; Imagery Intelligence or IMINT; Convolutional Neural Network or CNN; Computer Vision or CV; Autonomous Drones; Unmanned Aerial Vehicles or UAVs; Object Detection; Data Collection.	Artificial Intelligence or AI; Adversarial AI; Adversarial Attack; Poisoning Attack; Natural Language Processing or NLP; Data Preprocessing; Data Translation; Error Detection; Data Cleaning; Data Transformation.	Artificial Intelligence or AI; Adversarial AI; Adversarial Attack; Dual Denial of Decision Attack; Generative AI or GenAI; Large Language Model or LLM; Scenario Building; Data Analysis; Intelligence Analysis; Course of Action; Human-in-the-loop; Human-AI Teaming.
Scopus search query*	TITLE-ABS-KEY ((“Adversarial AI”) OR (“Adversarial Attack”) OR (“Evasion Attack”)) AND ((“Imagery Intelligence”) OR (“IMINT”) OR (“Convolutional Neural Network”) OR (“Computer Vision”) OR (“Autonomous Drones”) OR (“Unmanned Aerial Vehicles”) OR (“Object Detection”)) AND PUBYEAR > 2019	TITLE-ABS-KEY ((“Adversarial AI”) OR (“Adversarial Attack”) OR (“Poisoning Attack”)) AND ((“Natural Language Processing”) OR (“NLP”) OR (“Data Preprocessing”) OR (“Data Translation”) OR (“Error Detection”) OR (“Data Cleaning”) OR (“Data Transformation”)) AND PUBYEAR > 2019	TITLE-ABS-KEY ((“Adversarial AI”) OR (“Adversarial Attack”) OR (“Dual Denial of Decision Attack”)) AND ((“Generative AI”) OR (“Large Language Model”) OR (“Scenario Building”) OR (“Intelligence Analysis”) OR (“Data Analysis”) OR (“Human-in-the-loop”) OR (“Human-AI Teaming”)) AND PUBYEAR > 2019

*Due to the large number of search results, the search was limited to only the title, abstract, and keywords