AUGUST 20, 2025

MASTER THESIS

ETHICAL OVERSIGHT UNDER FIRE: EVALUATING THE ETHICAL IMPLEMENTATION OF LAWS WITHIN THE ROYAL NETHERLANDS ARMY

KJELL SCHIPPER
RADBOUD UNIVERSITY
S1005559

Abstract

This thesis investigates the applicability of the Comprehensive Human Oversight Framework (CHOF) in addressing the ethical and accountability challenges posed by Lethal Autonomous Weapon Systems (LAWS) within the Royal Netherlands Army (RNLA). As militaries increasingly explore autonomous systems, ensuring ethical oversight, accountability, and compliance with international humanitarian law has become a pressing concern. Although CHOF offers a strong theoretical model, its practical effectiveness has not been tested in military settings. This study employs an interpretivist, qualitative case study design, drawing on three policy documents and eleven semi-structured interviews with Dutch military stakeholders across operational, legal, and technical domains. Findings demonstrate that while the RNLA embeds international humanitarian law and responsibility structures deeply into its organizational culture, LAWS expose three persistent vulnerabilities: the limited feasibility of robust ex ante Article 36 reviews, the delegation of authority in the targeting cycle where LAWS cannot reliably fulfil international humanitarian law requirements, and the erosion of accountability in post-use review processes. CHOF provides value as a diagnostic and structuring tool, mapping oversight gaps across governance, sociotechnical, and technical layers, but falls short in prescribing solutions to these problems. The Glassbox Framework offers partial reinforcement by translating abstract norms into observable, auditable system behaviours, yet remains limited when applied to epistemic norms such as distinction, proportionality, and necessity. Its reliance on reliability, predictability, and traceability underscores the need for opening the “black box” and stronger institutional processes, including TEVV (Testing, Evaluation, Verification, and Validation), robust data governance, and sustained collaboration with industry partners. Taken together, the study concludes that CHOF and Glassbox provide useful conceptual anchors for oversight but cannot independently resolve the ethical and operational dilemmas posed by LAWS. Their effective implementation requires not only technical mechanisms, but also institutional adaptation, continuous testing, and governance integration tailored to military practice. Future research should investigate how data congruence, scenario-based TEVV, and algorithmic transparency can enhance the reliability and explainability of AWS. At the same time, organizational and cultural adaptation must be explored to ensure that CHOF and Glassbox function effectively in military practice.

Table of Contents

List of Abbreviations

- **AAR** – After Action Review
- **AGI** – Artificial General Intelligence
- **AI** – Artificial Intelligence
- **AIV** – Advisory Council on International Affairs
- **ATR** – Automatic Target Recognition
- **AWS** – Autonomous Weapon Systems
- **BDA** – Battle Damage Assessment
- **CAVV** – Advisory Committee on International Law Issues
- **CCW** – Convention on Certain Conventional Weapons
- **CHOF** – Comprehensive Human Oversight Framework
- **COMMIT** – Command, Materiel and IT (Dutch defense procurement program)
- **EW** – Electronic Warfare
- **GPS** – Global Positioning System
- **ICRC** – International Committee of the Red Cross
- **IHL** – International Humanitarian Law
- **ISR** – Intelligence, Surveillance and Reconnaissance
- **JWT** – Just War Theory
- **KMAR** – Koninklijke Marechaussee (Royal Netherlands Marechaussee)
- **LAWS** – Lethal Autonomous Weapon Systems
- **LIDAR** – Light Detection and Ranging
- **MHC** – Meaningful Human Control
- **ML** – Machine Learning
- **NATO** – North Atlantic Treaty Organization
- **NDD** – Netherlands Defense Doctrine
- **ODD** – Operational Design Domain
- **RNLA** – Royal Netherlands Army
- **ROE** – Rules of Engagement
- **SLAM** – Simultaneous Localization and Mapping
- **TEVV** – Testing, Evaluation, Verification and Validation
- **TNO** – Netherlands Organization for Applied Scientific Research
- **TRL** – Technology Readiness Level
- **UAS** – Unmanned Aerial System
- **UGV** – Unmanned Ground Vehicle
- **UN** – United Nations

1 Introduction

Lethal Autonomous Weapon Systems (LAWS) represent a transformative leap in military technology. These systems, capable of independently selecting and engaging targets, promise to revolutionize modern warfare by enhancing precision, reducing risks to human soldiers, and improving operational efficiency (Ecemis Yilmaz, 2023; Hammond, 2014). Their development is advancing rapidly on a global scale, with major powers such as the United States, China, and Russia heavily investing in their capabilities (AIV & CAVV, 2021; Kreps, 2021). This arms race underscores the strategic imperative for NATO members, including the Netherlands, to modernize their defence forces to remain competitive in an increasingly complex geopolitical environment (AIV & CAVV, 2021). The Dutch military's vision for 2035 emphasizes the creation of a "smart, technologically advanced defence organization" that integrates artificial intelligence and autonomy across domains, ensuring operational adaptability and effectiveness (Defensie, 2020, 2023). However, the adoption of LAWS comes with profound ethical challenges. While these systems offer significant operational advantages, their use raises critical questions about accountability, human oversight, and compliance with International Humanitarian Law (IHL) (Blanchard & Taddeo, 2024). Currently, no standardized approach is utilized to ensure that LAWS deployment aligns with ethical principles or international frameworks (Kwik, 2022a, 2022b). Moreover, there are not even binding international laws specifically governing the use of LAWS, leaving a regulatory vacuum that exacerbates concerns about misuse and a lack of accountability (Chengeta, 2016; Wagner, 2014). In response to these challenges, several conceptual frameworks have been developed to provide ethical oversight for autonomous weapon systems. Among these, the Comprehensive Human Oversight Framework (CHOF) in combination with the Glassbox framework is particularly notable for its systematic approach to embedding human oversight and accountability throughout the lifecycle of LAWS. The CHOF provides theoretical guidance but has no use cases in practical military contexts (Verdiesen, Santoni de Sio, et al., 2021). This gap between theory and real-world application highlights the pressing need for actionable strategies, policies and procedures, to bridge these pressing divides and gaps. This research addresses the critical need for ethical oversight in the deployment of LAWS within the Royal Netherlands Army **(**RNLA**)** by taking the CHOF as its central analytical instrument.

The practical relevance of this research lies in its capacity to address critical and transformational challenges in the ethical deployment of LAWS within the RNLA while upholding to broader international standards. As the RNLA navigates the integration of these transformative technologies, this study offers practical insights and actionable recommendations to ensure alignment with ethical and legal principles. Firstly, this research will help understand how military innovation in LAWS can coexist with strong ethical oversight as it studies the gap between theoretical ethical principles and their application in real-

world military contexts. This research will put the CHOF to the test in practical settings, examining how it can be integrated into everyday military decision-making processes. By providing concrete examples and strategies for embedding ethical considerations into LAWS operations, the study will serve as a vital resource for translating abstract principles into actionable practices. Additionally, as a member of NATO and a signatory to various international agreements, the Netherlands can set an international example for the ethical deployment of LAWS (AIV & CAVV, 2021). Secondly, the research directly aims to enhance Dutch military practices by identifying gaps in current approaches to ethical oversight and usage of LAWS. By systematically evaluating existing policies, procedures, and operational frameworks, this study will pinpoint areas requiring improvement and propose tailored enhancements to strengthen compliance with ethical war frameworks and IHL. These recommendations will not only help the RNLA refine its ethical oversight mechanisms but also ensure that LAWS deployment adheres to IHL. Finally, robust ethical oversight is essential for preventing potential failures in LAWS deployment. Without clear and enforceable ethical guidelines, LAWS risks undermining public trust, escalating conflicts, or violating international law (Dresp-Langley, 2023). Such failures could damage the legitimacy of military actions and jeopardize the protection of human rights (Margulies, 2017).

This research holds significant scientific value by addressing gaps in the academic discourse discussing the ethics of LAWS and the advancing in the development of actionable ethical frameworks. Firstly, this study evaluates the practical applicability of the CHOF. While frameworks, such as the CHOF, offer theoretical guidance on ethical oversight, their effectiveness remains largely untested, as there are no use cases in real-world military operations (Verdiesen, Santoni de Sio, et al., 2021). By applying this framework to the RNLA, this research identifies their strengths, limitations, and areas requiring refinement. Such empirical testing is critical for advancing these frameworks beyond the conceptual stage, making them more robust and suitable for operational use. A key contribution of this research lies in bridging the gap between ethical theory and military practice. There is a substantial disconnect between the philosophical debates on LAWS ethics and their application in practical settings. Current discussions often focus on abstract principles without providing clear guidance for their implementation (Ekelhof, 2019; Kwik, 2022a, 2022b). This study directly addresses this gap by situating theoretical insights within the specific context of the RNLA's operational environment. By doing so, it advances the field of applied military ethics, offering valuable lessons for the body of knowledge. Finally, this research addresses a critical gap in existing scholarship by focusing on the underexplored context of specific national militaries. Most academic studies on LAWS ethics concentrate on broad international or theoretical dimensions, often neglecting how these debates translate into practical decision-making at the national level (Kwik, 2022a). By examining how ethical frameworks are implemented or are planned to be implemented within the RNLA, this study provides a unique and valuable perspective that complements

the broader academic discourse. Given these gaps between ethical theory and military practice, this thesis positions the CHOF as its primary analytical lens for assessing how meaningful human control and accountability can be preserved in the deployment of autonomous weapon systems. Because CHOF's practical adequacy is uncertain, the Glassbox Framework is considered as a complementary approach that shifts assurance from internal model transparency toward observable, auditable behaviour. To examine the adequacy and adaptability of these frameworks within the RNLA the study is guided by the following question and sub-questions.

To what extent can the Comprehensive Human Oversight Framework, supplemented by the Glassbox Framework, be applied and adapted to ensure ethical, accountable, and operationally effective deployment of Lethal Autonomous Weapon Systems in the Royal Netherlands Army?

SQ1: What are the principal ethical challenges posed by Lethal Autonomous Weapon Systems, and how are these challenges reflected in the Royal Netherlands Army's current oversight practices?

SQ2: How do the ethical challenges identified in SQ1 expose weaknesses or gaps within the CHOF's layers and phases of oversight?

SQ3: To what extent can the integration of the Glassbox Framework with the Comprehensive Human Oversight Framework address the identified ethical and operational gaps in AWS oversight?

Together, these questions establish a stepwise logic: SQ1 identifies the ethical problem-set in practice; SQ2 diagnoses where, within CHOF's governance, socio-technical, and technical layers across pre-deployment, deployment, and post-deployment phases, control is weakened or lost; SQ3 then evaluates whether, and under what conditions, Glassbox can substantively reinforce CHOF to yield ethically meaningful and operationally viable oversight within the RNLA. This study is motivated by the urgent need to address the ethical dilemmas posed by LAWS and the lack of practical validation for existing frameworks. By examining the RNLA's practices considering the Comprehensive Human Oversight Framework, it seeks to address two critical gaps: the incomplete nature of current ethical frameworks and the absence of real-world operational testing. The findings aim to enhance the alignment of theoretical principles with military practices, advancing both academic understanding and the practical implementation of ethical guidelines for LAWS.

2 Literature Review

2.1 Defining Autonomous Weapon Systems

The definition of Lethal Autonomous Weapon Systems (LAWS) remains deeply contested, reflecting the divergent legal, ethical, and strategic interests of states, international organizations, and scholars (Taddeo & Blanchard, 2022). This definitional ambiguity has far-reaching implications: it shapes governance structures, the development of regulatory frameworks, and the ethical acceptability of autonomous weapons. Moreover, the act of defining LAWS is itself a political process, with stakeholders often leveraging definitions to promote particular regulatory objectives or technological interests (Taddeo & Blanchard, 2022). For the purposes of this thesis, I adopt the value-neutral definition articulated by Blanchard and Taddeo (2022). This approach is particularly well suited for empirical investigation, as it systematically incorporates the principal attributes identified in the literature while avoiding the normative constraints embedded in policy-driven definitions. According to Blanchard and Taddeo, an Autonomous Weapon System is:

"An artificial agent which, at the very minimum, is able to change its own internal states to achieve a given goal, or set of goals, within its dynamic operating environment and without the direct intervention of another agent, and may also be endowed with some abilities for changing its own transition rules without the intervention of another agent, and which is deployed with the purpose of exerting kinetic force against a physical entity (whether an object or a human being) and to this end is able to identify, select or attack the target without the intervention of another agent. Once deployed, AWS can be operated with or without some forms of human control (in, on, or out the loop). A lethal AWS is a specific subset of an AWS with the goal of exerting kinetic force against human beings." (Taddeo & Blanchard, 2022, p. 15)

This value-neutral approach provides conceptual clarity and analytical generalizability for the present study, while remaining compatible with the empirical realities of diverse military and policy contexts. It is, however, important to situate this analytical choice in relation to the Dutch national definition, which is frequently invoked in the policy and legal context of the Netherlands. The Dutch definition, as used by Verdiesen (2021) and formulated by the Advisory Council on International Affairs (AIV) and the Advisory Committee on International Law Issues (CAVV), defines an AWS as:

"A weapon that, without human intervention, selects and engages targets matching certain predefined criteria, following a human decision to deploy the weapon on the understanding that an attack, once launched, cannot be stopped by human intervention." (AIV & CAVV, 2015, p. 11)

Unlike the value-neutral approach, the Dutch definition embeds normative constraints, explicitly emphasizing human control by predefining targeting criteria. While these definitions differ in scope, the value-neutral approach can be seen as encompassing the Dutch definition, as it does not exclude human control but instead allows for a spectrum of regulatory models, including those mandating strict human control. In this sense, the Dutch definition represents a policy-driven refinement within the broader analytical framework adopted here. This distinction is crucial, as it enables the research to remain analytically comprehensive while maintaining direct relevance to the Dutch military context. Building on this definitional foundation, the next sections systematically analyse the four principal characteristics of LAWS as established by Taddeo and Blanchard (2022): autonomy, adapting capabilities, human control, and purpose of use. These dimensions structure the ensuing normative and empirical analysis.

Autonomy is a central concept in the discourse on AWS, raising significant legal, ethical, and operational concerns. The distinction between automated and autonomous systems is particularly crucial. The International Committee of the Red Cross (2019) emphasizes that automated systems operate based on pre-defined responses to environmental stimuli, whereas autonomous systems possess the ability to make decisions without direct human intervention. This differentiation is essential in understanding the varying degrees of autonomy within AWS, as the latter can adjust their behaviour in response to unpredictable environmental conditions. A more granular understanding of autonomy is provided by Castelfranchi and Falcone (2003), who classify autonomy into three distinct dimensions: executive autonomy, goal autonomy, and social autonomy. Executive autonomy refers to an agent’s ability to execute tasks without requiring real-time human input, a characteristic exhibited by many AWS in their targeting and engagement functions. Goal autonomy extends this capability by allowing systems to independently determine their objectives based on situational awareness, thereby increasing the complexity and unpredictability of their actions. Social autonomy, in contrast, relates to an agent’s ability to interact with other agents, whether human or artificial, and modify its behaviour accordingly. This aspect of autonomy raises concerns about AWS operating in dynamic, multi-agent environments, where unintended escalation could occur due to unanticipated interactions with other systems. Floridi and Sanders (2004) contribute further to the conceptualization of autonomy by identifying three key attributes necessary for an entity to be considered autonomous: interactivity, autonomy, and adaptability. Interactivity refers to the system's capacity to perceive and act upon its environment dynamically. AWS with high levels of interactivity can adjust their operational parameters in response to real-time threats, which enhances their effectiveness but simultaneously reduces predictability. Autonomy, in this framework, is defined as an agent's ability to modify its internal states without direct external intervention. This characteristic distinguishes AWS from purely automated systems, as they can alter their decision-making strategies based on internal processes rather than fixed programming. Lastly, adaptability refers to the system's ability to refine its behaviour

over time, often through machine learning mechanisms. Adaptability is defined as a different category by Blanchard and Taddeo (2022), and will be discussed in the next section.

Adapting capabilities refer to the ability of AWS to modify their behaviour in response to environmental changes without direct human intervention. These capabilities often rely on Artificial Intelligence (AI) techniques, particularly machine learning, to enhance the system's performance beyond pre-programmed responses. Machine learning plays a fundamental role in enabling AWS to develop adaptive capabilities. The International Committee of the Red Cross (ICRC, 2019) defines machine learning as a technique that allows AI systems to improve performance through exposure to data rather than relying solely on fixed programming. This learning process can take two primary forms: supervised learning, in which AI is trained on labelled datasets, and unsupervised learning, where the system autonomously identifies patterns in data. Traditionally, machine learning consists of two stages: training, where the system learns from vast amounts of data, and deployment, where it applies that learning in real-world contexts. These two steps are usually kept separate in most civilian applications, ensuring that training stops before the algorithm is deployed, a method known as off-line learning. However, some algorithms continue learning after deployment, a process known as online learning, in which the system constantly updates and modifies its decision-making model based on new data inputs (ICRC, 2019).The unpredictability of machine learning systems stems from their capacity to evolve beyond their original programming, making it difficult for operators to foresee how AWS will behave in novel combat scenarios (ICRC, 2019). This raises concerns about the ability of these systems to comply with legal and ethical constraints, particularly in situations requiring proportionality and distinction in targeting. Bode and Watts (2023) further emphasize the implications of machine learning in AWS, particularly in the context of loitering munitions. They highlight that machine learning enables these systems to engage in real-time adaptation, refining their targeting algorithms based on previous interactions. However, this introduces a critical issue: the potential for AWS to develop behaviours that were not anticipated or tested before deployment. This "black box" effect, where the decision-making process becomes opaque even to developers, increases the risk of unlawful or unethical engagements. Furthermore, as AWS gain greater autonomy in learning and adapting, the role of human oversight diminishes, making it more challenging to attribute responsibility for their actions. The significance of adapting capabilities in defining AWS cannot be overstated as they impact the predictability and controllability of AWS, complicating efforts for ethical control.

The ethical concern surrounding adaptive AWS is closely linked to the principle human control. The ICRC (2019) warns that as machine learning increases unpredictability; human operators may struggle to intervene effectively when AWS make inappropriate targeting decisions. This risk is particularly concerning in environments where AWS are deployed for extended periods with minimal human

oversight. Human Control is described as the third category of Blanchard and Taddeo (2022). Human control is a central element in the ethical and legal discourse surrounding AWS. While autonomy in AWS enables them to operate independently from human intervention, Taddeo and Blanchard (2022) emphasize that this autonomy does not negate the necessity of human oversight. They highlight that human control and AWS autonomy are not inherently contradictory but can coexist across different levels, ranging from political and strategic decision-making to the specific operational tasks delegated to these systems. The key issue, they argue, is not whether human control should exist, but rather which forms of control are ethically desirable and should inform governance frameworks (Taddeo & Blanchard, 2022). Their analysis underscores that discussions on human control should go beyond mere oversight and focus on defining the appropriate mechanisms to ensure compliance with IHL and ethical principles. Amoroso and Tamburrini (2021) identify three fundamental forms of human control that must be embedded in AWS operations: fail-safe control, which ensures that system malfunctions do not result in excessive or unlawful harm; accountability control, which establishes clear responsibility for AWS actions; and moral agency control, which guarantees that life-and-death decisions adhere to human ethical and moral frameworks. These three dimensions are crucial in addressing the governance challenges AWS pose, particularly as machine-learning capabilities introduce new risks of unpredictability and adaptation beyond initial programming (ICRC, 2019). The necessity of human control in AWS definitions extends beyond operational safety, it is a prerequisite for ethical and legal governance. Without meaningful human control, AWS risk being deployed in ways that undermine accountability, increase the likelihood of unintended engagements, and blur legal responsibility in conflict settings. While this report does not delve into the normative conditions for AWS design, development, and deployment, as noted by Taddeo and Blanchard (2022), it is crucial to recognize that human control must be incorporated at multiple levels to maintain compliance with existing legal and ethical norms.

The purpose of use is a fundamental aspect in defining AWS, as it directly relates to the intended function and ethical implications of these technologies. Taddeo and Blanchard (2022) highlight that most definitions of AWS implicitly qualify their purpose through references to “weapons” and their deployment in kinetic contexts. This framing suggests a destructive purpose, whether anti-material or lethal, but also reveals gaps in defining the specific tasks AWS perform within combat operations. The absence of clear distinctions in AWS definitions has led to ambiguities regarding whether these systems are primarily used for deliberate or dynamic targeting, and how they integrate into broader strategic and operational frameworks. The purpose of deployment is inherently linked to the objective an AWS is designed to achieve. Taddeo et al. (2022) argue that AWS must be evaluated based on their intended military function, as their deployment aligns directly with the goals of a given operation. The term Lethal Autonomous Weapon Systems (LAWS) derives from this lethal application of AWS. The LAWS

designation specifically pertains to systems designed to exert kinetic force against human targets, whereas AWS more broadly encompass systems with varying levels of autonomy and diverse operational goals. The distinction between AWS and LAWS carries significant ethical and legal implications, particularly regarding compliance with IHL and discussions on prohibitions or restrictions on autonomous weaponry. In this thesis AWS and LAWS will be used interchangeably, depending on the context.

2.2 Significance and functions of AWS

Having established the conceptual foundations and definitional contours of AWS, the discussion now turns to their broader significance within contemporary military and strategic contexts. Understanding the role and impact of AWS extends beyond their technical attributes; it also requires attention to the transformative implications these systems hold for defence organizations, operational effectiveness, and the ethical and legal frameworks governing the use of force. Strategically, the defining characteristics of autonomy and adaptability make AWS highly attractive to militaries aiming to maintain or enhance their operational advantage. The capacity for independent decision-making and real-time adjustment on the battlefield addresses the growing need for flexibility and rapid response in increasingly complex and dynamic security environments (Altmann & Sauer, 2017; Verbruggen & Boulanin, 2017a). States such as the United States, China, and Russia have prioritized these capabilities in their defence modernization programs, viewing AWS as essential for sustaining military competitiveness in an era marked by both conventional and hybrid threats (Altmann & Sauer, 2017). The autonomy inherent in AWS enables them to execute critical functions direct human input, thereby reducing decision latency and enabling faster, more coordinated military actions. Operationally, the adaptive capabilities afforded by machine learning are revolutionizing traditional military processes. AWS can process vast streams of sensor data, adjust their behaviour based on environmental changes, and perform tasks that would otherwise demand significant human resources and cognitive load. This not only enhances mission effectiveness but also reduces the exposure of human soldiers to direct risks (Verbruggen & Boulanin, 2017a). For example, AWS are increasingly used in high-risk or high-tempo operations where human intervention would be impractical or unsafe. Their ability to maintain operational tempo under complex and contested conditions further underscores their strategic value. Economically, the automation of core military functions made possible by AWS addresses persistent budgetary pressures and manpower constraints. Although the initial costs associated with the development and integration of AWS are substantial, their long-term use is often justified by projected cost savings through force multiplication and decreased reliance on human personnel (Verbruggen & Boulanin, 2017a). AWS can operate continuously with minimal human oversight, allowing for more efficient allocation of human resources and the possibility of maintaining larger operational footprints without proportional increases in personnel. Technological innovation,

particularly advances in artificial intelligence and sensor fusion, is another major driver of AWS adoption. The rapid pace of civilian technological advancement has led to significant “spillover” effects in the defence sector, as dual-use AI and robotics technologies are increasingly adapted for military applications (Altmann & Sauer, 2017). This integration supports the ongoing evolution of AWS capabilities, making it possible to endow these systems with higher levels of autonomy, adaptability, and mission-specific functionality. The result is an accelerating cycle of technological innovation and operational transformation, wherein the four principal characteristics of AWS, autonomy, adaptive capability, human control, and purpose, serve as both the rationale for and the mechanism of their growing significance within contemporary armed forces. To understand how these abstract drivers materialize in practice, the following section examines the core functions of autonomy in AWS, illustrating how these features are operationalized on the battlefield and what implications they hold for military effectiveness and oversight.

The most important autonomous function within weapon systems, and the primary focus of this research, is targeting: the system’s capacity to detect, identify, classify, and, in some cases, select or prioritize targets without direct human input. This function, often implemented through Automatic Target Recognition (ATR), raises the most acute ethical and legal concerns. As Verbruggen and Boulanin (2017) explain, ATR systems operate by comparing sensor inputs, such as radar, thermal, or visual imagery, against a library of predefined templates or profiles to identify objects like vehicles, aircraft, or personnel. These systems function on the basis of pattern recognition, enabling platforms to detect targets even under degraded visual conditions. However, their capabilities are usually narrow and task-specific, with limited adaptability to novel scenarios or unfamiliar target types.

Modern targeting autonomy is increasingly driven by advances in computer vision and machine learning. Computer vision allows AWS to process and interpret visual data from onboard cameras or sensors, identifying spatial features such as human forms, weapons, or military hardware in real time (Hughes, 2020; ICRC, 2019). This is often paired with deep learning models, particularly convolutional neural networks (CNNs), that are trained on large datasets to improve object classification accuracy and reduce false positives (Hughes, 2020; Kwik, 2024). Despite these advancements, most ATR systems remain static post-deployment, meaning they cannot learn or adapt to new target profiles during use, referred to as offline-learning (Verbruggen & Boulanin, 2017a). Some systems also incorporate target prioritization, using rule-based logic to rank targets based on mission criteria such as threat level, proximity, or strategic value. However, this process remains highly dependent on human-programmed parameters, which limits the system’s autonomy in more complex, ambiguous combat scenarios. As Bode and Watts (2023) argue, while such targeting technologies can increase speed and reduce operator workload, their lack of contextual understanding remains a major technical constraint. Nonetheless, targeting remains one of the

most strategically significant autonomous functions, as it directly shapes how force is applied on the battlefield and is the core site for ethical contestation over AWS.

Alongside targeting, navigation is another core autonomous function in weapon systems, enabling platforms to independently move through their environment and adapt to changing operational conditions (Verbruggen & Boulanin, 2017a). Unlike manually piloted or remotely controlled systems, autonomous navigation enables a platform to execute tasks such as take-off, route traversal, landing, and last mile targeting without human intervention. Early systems relied on pre-programmed waypoint navigation, where a system follows a set of GPS coordinates. However, more advanced platforms, such as the MQ-4C Triton and X-47B, demonstrate capabilities for autonomous take-off and landing, using onboard sensors to assess runway conditions and calibrate descent and lift-off trajectories in real time (Verbruggen & Boulanin, 2017a). These functions reduce dependency on communication links and enhance operational flexibility. A key enabler of this autonomy is sensor fusion, which integrates data from multiple sources, such as radar, LIDAR, inertial measurement units (IMUs), and visual cameras, to create a coherent understanding of the environment (Longpre et al., 2022). This allows systems to perceive obstacles, determine their own position when GPS is denied, and continuously update their path. Machine learning algorithms further enhance this process by learning from environmental data to improve decision-making over time. For instance, reinforcement learning enables systems to refine routing strategies based on trial-and-error in simulated or real-world environments (ICRC, 2019; Kwik, 2024). Moreover, computer vision technologies are critical for real-time environmental assessment. These systems interpret visual inputs, such as terrain features, buildings, or other objects, to support dynamic routing and ensure safe manoeuvring (ICRC, 2019). Ukraine's use of drones that navigate the "last mile" autonomously demonstrates how these integrated technologies are being applied in practice, allowing drones to locate, approach, and strike targets with minimal operator input (Bondar, 2025).

The intelligence function in AWS refers to the autonomous collection, processing, and analysis of environmental or operational data to support decision-making and mission execution. Unlike targeting, which focuses on object identification, intelligence autonomy enables broader tasks such as threat assessment, terrain mapping, and pattern recognition across communication or movement data (Verbruggen & Boulanin, 2017a). Systems like Shield AI's Nova exemplify this by autonomously generating 3D maps of indoor environments using onboard sensors and Simultaneous Localization and Mapping (SLAM) techniques (Longpre et al., 2022). These capabilities are particularly valuable in GPS-denied areas and urban warfare settings. Central to this function is machine learning, which enables systems to extract meaning from large, unstructured datasets. For example, machine learning models can analyse drone surveillance footage or intercepted communications to infer enemy locations or predict

movements (Bondar, 2025). Sensor fusion further enhances intelligence autonomy by integrating inputs from multiple modalities, visual, audio, infrared, into a coherent situational picture (ICRC, 2019). Though these systems typically do not initiate strikes autonomously, their analytical role significantly enhances operational tempo and situational awareness. As AWS evolve, intelligence functions increasingly blur the line between passive observation and proactive mission support. The embeddedness of intelligence functions in AWS, and information shared to larger decision models implies the dangers of implementing AWS in larger command and control structures.

The interoperability function of AWS refers to their ability to operate in coordination with other systems, whether human-operated or autonomous, through the sharing and interpretation of data. What distinguishes modern AI-enabled interoperability from earlier, rule-based systems are its capacity for dynamic adaptation. Traditional systems relied on pre-programmed protocols, functioning only under known conditions and following rigid instructions. In contrast, AI-enabled systems, particularly those using machine learning and computer vision, can interpret diverse inputs and make context-sensitive decisions in real time (Kwik, 2024). In machine-to-machine (M2M) coordination, AI allows systems to go beyond merely exchanging data: they can evaluate its relevance, extract patterns, and autonomously update their behaviour. For example, a reconnaissance drone using machine learning can detect enemy movement, flag unusual activity, and relay information directly to a loitering munition, which then adjusts its holding pattern or alert posture accordingly (Verbruggen & Boulanin, 2017a). This level of interoperability would be infeasible in conventional systems lacking adaptive learning algorithms. Similarly, in machine-to-human (M2H) coordination, AI technologies like computer vision allow systems to recognize battlefield features, classify threats, and generate visual or textual outputs that operators can understand and act on. Rather than requiring step-by-step commands, operators can issue broad objectives, trusting the system to interpret and execute tasks based on situational awareness. Longpre et al. (2022) emphasize that sensor fusion, integrating radar, infrared, video, and geolocation data, further enhances shared understanding across platforms. A particularly promising application of interoperability is swarming: the deployment of large numbers of autonomous agents acting in coordination. Swarms can distribute tasks, adapt formations, and maintain collective behaviour without centralized control, offering unparalleled flexibility, resilience, and operational tempo. While still emerging, swarming represents the most advanced form of interoperability, and a transformative frontier for autonomous warfare.

The health management function in AWS enables platforms to monitor, assess, and respond to their own internal states without human input. This includes real-time evaluation of mechanical integrity, sensor performance, battery levels, and system malfunctions. Unlike traditional systems that rely on operator diagnostics or scheduled maintenance, machine learning powered AWS can use predictive analytics,

sensor fusion, and anomaly detection algorithms to identify faults early and adapt behaviour accordingly (Kwik, 2024). AI is vital here because it allows for continuous self-assessment in dynamic and remote environments, reducing downtime, enhancing resilience, and enabling longer, more autonomous missions with minimal logistical support.

While these autonomous functions, especially targeting, are often discussed in theoretical or experimental terms, their operational significance has become starkly apparent in recent conflicts. The ongoing war in Ukraine, in particular, provides a concrete and urgent illustration of how AWS are rapidly shifting the realities of modern warfare. Both Ukrainian and Russian forces have deployed a wide range of systems with autonomous capabilities, fundamentally reshaping combat dynamics (Bondar, 2025; Rickli & Mantellassi, 2024). Ukraine, facing significant manpower shortages and operational pressures, has heavily invested in modular, AI-enabled drones that can autonomously navigate complex environments, identify targets, and, in some cases, deliver lethal strikes without direct human intervention (Bondar, 2025; Kunertova, 2024). Russian forces, likewise, have employed loitering munitions such as the Lancet and KUB-BLA, which patrol battlefields autonomously and engage targets based on algorithmic recognition (Blakcori et al., 2024; Saxon, 2024). Across the frontline, the saturation of low-cost, semi-autonomous drones has become a defining feature of operations, with both sides increasingly using these systems not only for reconnaissance but for offensive engagements (Kunertova, 2024; Molloy, 2024). One of the clearest indicators of this shift is the implementation of ATR systems in Ukrainian drone warfare, in the form of last-mile navigation. Without ATR, drone strikes relied heavily on manual targeting, with operators guiding drones under conditions vulnerable to human error, fatigue, and electronic warfare interference (Molloy, 2024). Strike success rates under manual operation were relatively low, often below 20% in contested environments (Bondar, 2025). However, with the integration of ATR, which allows drones to autonomously help army personnel detect, track, and prioritize targets, Ukrainian forces have achieved success rates between 70 and 80 percent (Bondar, 2025). This operational effectiveness highlights the tactical advantages of autonomy, enabling faster engagements, greater resilience under jamming conditions, and reduced operator workload (Kunertova, 2024; Rickli & Mantellassi, 2024). Yet this same effectiveness underscores the dangers of autonomy: once ATR systems are activated, critical decisions about life and death are made by machine learning algorithms at machine speed, leaving no space for real-time human ethical judgment (Saxon, 2024). Electronic warfare can cause lost connections, preventing any monitoring of strike success. These trends reveal a dangerous erosion of the traditional model of human-centred warfare. As autonomy advances, human operators are relegated to supervisory roles, often involved only in the initial deployment decision, while navigation, targeting, and engagement unfold autonomously (King, 2024; Kunertova & Herzog, 2024). The battlefield experience in Ukraine shows that meaningful human control is often more theoretical than real, with human intervention nearly

impossible once autonomous systems are in motion (Bondar, 2025; Kunertova, 2024). Moreover, neither Ukraine nor Russia has implemented clear legal or ethical frameworks governing AWS-like systems, resulting in serious accountability gaps and challenges to compliance with international humanitarian law principles such as distinction, proportionality, and necessity (Kunertova & Herzog, 2024; Saxon, 2024). The result is a battlefield environment characterized by heightened unpredictability, blurred responsibility, and a widening gap between military effectiveness and ethical governance (Kunertova & Herzog, 2024; Rickli & Mantellassi, 2024). Instead of reducing the uncertainties of war, autonomous systems have deepened the fog of war, making attribution for unlawful acts increasingly difficult and allowing ethical violations to occur without clear chains of accountability(Saxon, 2024). For the RNLA, the implications are profound. Future conflicts will likely involve adversaries deploying autonomous and semi-autonomous systems in large volumes, overwhelming traditional command structures and decision-making cycles (Blakcori et al., 2024; King, 2024). Without urgently embedding robust frameworks for ethical oversight, meaningful human control, and transparent accountability into future military technologies, the Dutch military risks serious operational vulnerabilities and legal and moral exposure. The Ukrainian experience stands as a clear warning: AWS are not a future problem, they are already reshaping warfare, and militaries that fail to adapt to their operational and ethical challenges will find themselves at a critical disadvantage.

As demonstrated in Ukraine, the promises of enhanced efficiency, force projection, and reduced human exposure are inseparable from the significant ethical and legal dilemmas introduced by increased autonomy and adaptive capacity (ICRC, 2019; Taddeo & Blanchard, 2022). The unpredictability inherent in machine learning complicates core International Humanitarian Law principles such as proportionality and distinction, while the erosion of meaningful human control in real-world combat raises serious concerns regarding accountability and ethical governance (Bode & Watts, 2023; Poitras, 2018). Although some proponents claim that AWS can, in theory, outperform humans in legal compliance, these assertions remain highly contested and, as current conflicts reveal, often unsubstantiated in practice (Riesen, 2022). As AWS continue to evolve beyond the limits of direct human oversight, they risk undermining foundational principles of ethical warfare and legal responsibility (Asaro, 2012; Chengeta, 2016). The next chapter will therefore critically examine these dilemmas, focusing on the ethical implications of autonomous decision-making in armed conflict and the potential for AWS to fundamentally dehumanize the use of force.

2.3 Ethics and International Humanitarian Law

The ethical regulation of warfare has a long tradition rooted in moral philosophy, theology, and later international law. Just War Theory emerged to reconcile the necessity of war with ethical constraints, first articulated by Augustine and Aquinas, and later secularized by Grotius into principles of natural law applicable to sovereign states (Seixas-Nunes, 2020; Walzer, 2015). It developed into two core branches: *jus ad bellum*, outlining legitimate reasons for war, and *jus in bello*, regulating conduct during warfare through principles such as distinction, proportionality, necessity and humanity (Blauth, 2023). These moral principles were gradually codified into IHL, culminating in the Geneva Conventions of 1949 and their Additional Protocols, which remain the cornerstone of modern law of armed conflict (ICRC, 2011). IHL requires states to ensure new weapons comply with these norms, notably through Article 36 reviews, and upholds fundamental principles: distinction between civilians and combatants, proportionality in force, and necessity to achieve legitimate military objectives (Christie et al., 2024). However, both Just War Theory and IHL have limitations. They focus on the morality and legality of entering and conducting war but cannot address underlying political causes or ensure compliance. This gap underscores the need to complement legal norms with continuous ethical reflection and political oversight, particularly in the context of emerging military technologies (Blauth, 2023).

The ethical and legal frameworks of JWT and IHL are not only foundational in philosophical and legal scholarship but also explicitly integrated into the operational doctrines of modern states. The Dutch Defence Doctrine (NDD) reflects this by emphasizing the importance of the “morele component” (moral component) of military power, which underscores that military operations must be conducted with both legal and moral legitimacy (Defensie, 2025, p. 15). The NDD adopts the dual structure of jus ad bellum and jus in bello, noting that the legality of the use of force is determined by international law, particularly the UN Charter, and that the conduct of warfare must adhere to the principles of IHL, as codified in the Geneva Conventions and their Additional Protocols (Defensie, 2025, pp. 18-19). The NDD specifically recognizes that IHL seeks to achieve a balance between the necessity of military action and humanity, by limiting the means and methods of warfare and protecting persons and objects affected by conflict. It stresses that these rules apply in all cases of armed conflict and, even when IHL does not formally apply, the Netherlands and NATO choose to observe its protective provisions to maintain moral legitimacy (Defensie, 2025, p. 19; NATO, 2024). Through this explicit incorporation of legal and ethical principles, the NDD demonstrates the enduring relevance of JWT and IHL in guiding the responsible use of military force. However, the rapid advancement and integration of AWS now place these foundational doctrines under unprecedented strain. While Dutch military doctrine insists on moral legitimacy and strict adherence to IHL, the delegation of critical functions to AWS raises pressing questions about whether

these standards can be upheld in practice. As AWS take on greater roles in targeting and engagement, they increasingly disrupt the core assumptions that underpin both legal and ethical accountability in warfare.

The deployment of AWS poses significant legal and ethical challenges to the core principles of IHL discussed before and thus pose a risk to the RNLA's doctrines. Each of the foundational tenets, proportionality, necessity, and distinction, is grounded in the notion that humans exercise judgment under conditions of uncertainty and moral responsibility; capacities that AWS, by their very nature, do not inherently possess (Boutin & Woodcock, 2024; Hughes, 2020; Kwik, 2024). The principle of proportionality prohibits attacks that may cause excessive civilian harm in relation to the anticipated military advantage. Yet AWS, particularly those employing machine learning, lack the contextual awareness and ethical reasoning needed to reliably assess proportionality in dynamic combat environments (Boutin & Woodcock, 2024; Devitt, 2024). Unlike human commanders who can exercise discretion based on immediate situational awareness, AWS operate through pre-programmed rules or learned patterns, which may not account for rapidly evolving civilian presence or the nuanced value of military targets. Bode and Watts (2023) highlight that loitering munition using adaptive targeting algorithms risk recalibrating their strike decisions based on training data, potentially prioritizing effectiveness over proportionality. Moreover, as Schwarz (2021) argues, proportionality evaluations inherently require normative assessments, such as judging whether civilian casualties are "excessive", which cannot be reduced to quantifiable inputs without a significant loss of ethical nuance. Thus, delegating proportionality decisions to AWS shifts these judgments away from human moral agents toward opaque algorithmic processes.

Closely related is the principle of distinction, which mandates the differentiation between lawful military targets and protected civilians. While AWS may be equipped with ATR or computer vision systems, these technologies are prone to error, particularly in complex environments such as urban warfare where distinguishing combatants from non-combatants is often ambiguous (Christie et al., 2024; Kwik, 2024). The risk is exacerbated by the lack of contextual reasoning in AWS; a human soldier might infer intent based on subtle behavioural cues, whereas an AWS must rely on rigid classification schemas or probabilistic models (Booker, 2024). According to Ebrahimi (2024), reliance on such models introduces a margin of error that is ethically untenable when life-and-death decisions are at stake. Furthermore, AWS systems may operate under degraded conditions, such as GPS denial, sensor interference, or incomplete data, which further diminishes their capacity to uphold the principle of distinction (Boulanin & Lewis, 2023). While defenders of AWS argue that these systems may reduce emotional bias or fatigue-induced errors common in human operators, this utilitarian claim does not resolve the core issue that AWS lack the cognitive and empathetic faculties required to make moral distinctions (Taddeo & Blanchard, 2022).

The principle of necessity requires that any use of force be limited to what is essential to achieve a legitimate military objective. Traditionally, this has meant a requirement to avoid excessive or redundant violence, and to pursue less harmful alternatives when available (Christie et al, 2024). However, the operational autonomy of AWS, especially those operating beyond human supervision in "out-of-the-loop" modes, undermines the assurance that lethal actions are truly necessary (Kwik, 2024). For example, Devitt (2024) notes that AWS may engage targets pre-emptively based on probabilistic threat modelling, rather than real-time situational judgments that a human operator would employ. This can result in the use of force in scenarios where de-escalation or retreat might have been a viable alternative (Devitt, 2024). Additionally, the coding of necessity into AWS involves the reduction of ethical judgments to computational thresholds, such as defining "imminent threat" based on pattern recognition, which risks expanding the use of force beyond what would be considered necessary under IHL (Amoroso & Tamburrini, 2021). The problem is not only technical but conceptual: machine agents cannot grasp the human values embedded in the concept of military necessity, which has traditionally relied on a balancing of strategic goals against humanitarian concerns (Blauth, 2023, Hughes, 2019).

The principle of humanity, which prohibits means and methods of warfare that cause unnecessary suffering, is arguably the most philosophically contested in the context of AWS. According to Boulanin and Lewis (2023), the use of AWS may reduce certain forms of suffering, such as physical exhaustion or psychological trauma experienced by human soldiers, but it simultaneously risks dehumanizing warfare by severing the link between human empathy and the exercise of lethal force. Schwarz (2021) emphasizes that the absence of human agency in AWS decisions erodes the moral gravity of killing, transforming ethical deliberation into a technical process devoid of compassion or remorse. Furthermore, the deployment of AWS may undermine humanitarian obligations not because they are less capable of precision, but because they decouple killing from human moral intuition (Ebrahimi, 2024; Schwarz, 2021). Oimann and Tollon (2025) critique the assumption that reducing human suffering on one side of the conflict justifies a technological means that may increase suffering on the other, noting that humanity must be assessed from a universal ethical standpoint, not just from the perspective of military efficiency.

Underlying these doctrinal challenges is the deeper ethical dilemma of delegating moral decision-making to machines. The question is not simply whether AWS can technically perform tasks currently undertaken by humans, but whether they ought to be entrusted with decisions that have life-or-death consequences. As Santoni de Sio and Van den Hoven (2018) argue, moral responsibility requires agency, foresight, and the capacity to justify one's actions within a normative framework, all characteristics absent in machine agents. When AWS are used to make targeting decisions, the chain of moral accountability is disrupted, as there is no clear subject to which moral blame or praise can be ascribed (Schwarz, 2021). This erosion of

human agency not only affects post-strike accountability but also weakens pre-strike ethical deliberation, as operators may defer judgment to the "objectivity" of the system (Umbrello, 2021). Consequently, moral agency in warfare risks being diluted by an overreliance on technological delegation.

The legal implications of this ethical delegation are most apparent in the context of Article 36 weapons reviews. These reviews require states to assess whether new weapons, means, or methods of warfare comply with IHL before deployment. However, adaptive AWS complicate this process. Because such systems may change behaviour post-deployment, particularly those using online machine learning, their full operational impact cannot be accurately assessed in a one-time pre-deployment review (Christie et al., 2024; Poitras, 2018). Ongoing learning introduces a "moving target" problem: even if a system passes legal review at one point in time, it may evolve to act unlawfully later without operator awareness (Verbruggen & Boulanin, 2017b). Moreover, there is currently no consensus among states on how to evaluate AWS under Article 36, leading to fragmented and inconsistent application of legal standards (Cavalcante Siebert et al., 2023). Without robust procedures for continuous monitoring and re-assessment, Article 36 becomes an inadequate safeguard against unlawful behaviour by adaptive AWS.

Adding to these challenges is the absence of specific international laws regulating AWS. While some international instruments, such as the Convention on Certain Conventional Weapons (CCW), have hosted discussions on AWS, no binding treaty currently exists that directly addresses their development, deployment, or use (Ebrahimi, 2024). The definitional ambiguity of AWS, ranging from semi-autonomous systems to fully independent agents, has further hindered regulatory consensus (Blanchard & Taddeo, 2022). This legal vacuum not only creates room for divergent national standards but also incentivizes strategic ambiguity and technological arms races among states. As Taddeo and Blanchard (2022) note, states may exploit this legal uncertainty to develop systems that skirt ethical responsibilities while maintaining plausible deniability. The lack of codified norms means that existing IHL principles must be stretched to accommodate a new class of actors whose actions and intentions do not fit traditional legal categories.

2.4 Accountability

When AWS violate core IHL principles such as distinction, proportionality, and necessity, fundamental questions emerge concerning accountability: Who is responsible when these systems cause unlawful harm? Conventional military accountability frameworks are rooted in clear human agency and defined chains of command, where decisions are traceable to identifiable individuals who can subsequently be held legally, morally, and institutionally accountable (Smith, 2022; Wood, 2023). However, the integration of AWS, particularly those with advanced machine-learning capabilities and minimal human

oversight, significantly disrupts these traditional accountability processes. Decisions made autonomously by these systems blur the distinction between human intention and machine action, complicating the assignment of responsibility and the effectiveness of legal and ethical evaluations when violations occur (Boulanin & Lewis, 2023; Cavalcante Siebert et al., 2023). In addressing these challenges, a clear understanding of accountability is required. While accountability and responsibility are often used interchangeably, they represent distinct normative concepts. Responsibility encompasses both forward-looking and backward-looking dimensions, addressing not only moral and causal ownership of actions and outcomes but also proactive obligations to prevent harm (Bovens, 2007; Verdiesen, Santoni de Sio, et al., 2021). Accountability, conversely, is exclusively backward-looking, referring specifically to institutionalized mechanisms through which actors explain and justify their past actions to relevant authorities such as courts, superiors, or the public, and face potential sanctions if necessary (Bovens, 2007). Given the focus of this research on ethical implementation within the RNLA, an organizational context with clearly institutionalized accountability procedures, this research will predominantly delve into the accountability dimension, as conceptualized by Bovens. Bovens (2007) conceptualizes accountability as a relational structure between an actor and a forum, comprising elements such as information provision, debate, judgment, and potential sanctions. Importantly, accountability as described by Bovens does not guarantee that genuine responsibility exists; actors can be held accountable even without being genuinely responsible. For example, a political minister might face consequences for departmental failures that occurred without direct personal fault. Bovens further categorizes accountability into five distinct modalities, political, legal, administrative, professional, and social, each functioning within clearly defined institutional relationships and responsibilities. This conceptual clarity allows for systematic examination of accountability within military structures deploying AWS.

The operational deployment of AWS, particularly those employing advanced adaptive capabilities, significantly undermines Bovens' relational model of accountability. The core issue arises from the dispersed decision-making characteristic of AWS deployments: multiple actors such as developers, commanders, system operators, and automated decision algorithms collectively participate without any single actor necessarily possessing complete oversight at the critical moment of action (Smith, 2022; Wood, 2023). Consequently, when AWS cause unlawful harm, there is no singular or clear chain of accountability. This introduces what has been termed an "accountability vacuum," where the absence of clear intentionality, moral perception, and human justification in AWS actions destabilizes traditional military accountability frameworks(Cavalcante Siebert et al., 2023; Ebrahimi, 2024; Kwik, 2022a; Umbrello, 2021). This accountability disruption is significantly intensified by the inherent opacity of AWS, especially those utilizing machine learning algorithms. These systems typically operate as "black

boxes," where internal decision processes remain inaccessible even to system developers or operators (Bode & Watts, 2023). Such opacity profoundly complicates both traceability, the ability to reconstruct the chain of events leading to a decision and connecting these to a human actor in the system, and explainability, the ability to rationalize why a decision occurred, critical elements for accountability investigations (Christie et al., 2024). The lack of these capacities results in insurmountable attribution dilemmas: when an AWS misclassifies a target, determining whether the error was due to training data, algorithmic inadequacy, operational misuse, or unforeseeable battlefield conditions becomes nearly impossible (Devitt, 2024). Without effective traceability and explainability mechanisms, legal and ethical oversight is fundamentally weakened. Courts and review bodies, lacking necessary evidentiary bases, are left either to trust system outputs blindly, a proposition incompatible with democratic norms, or to reject AWS altogether, limiting their operational use (Kwik, 2024). Thus, the absence of clear auditing pathways within AWS decisions undermines the foundational accountability structure of democratic military institutions like the RNLA (Christie et al., 2024).

This accountability does not only undermine the enforcement of legal norms but also erodes the perceived legitimacy of the military institution itself. Bovens (2007) highlights that accountability is not merely a procedural requirement but a democratic safeguard that ensures state institutions act within the boundaries of public trust and ethical justification. In his evaluative model, he outlines three key dimensions of accountability: democratic legitimacy, prevention of power abuse, and institutional learning. AWS threaten all three. First, the opacity of AWS operations impairs the ability of elected officials and citizens to meaningfully assess and control military force deployment, thereby compromising democratic accountability. Second, the inability to hold any actor definitively responsible for mistakes, such as misfires or civilian casualties, renders institutional safeguards against abuse of power ineffective. Third, the absence of traceable decision logs and explainable outputs limits the military's capacity for organizational learning, reducing opportunities to identify and rectify flaws in AWS deployment or targeting logic (Booker, 2024; Kwik, 2024). Thus, the introduction of AWS introduces systemic fragility to every level of accountability infrastructure envisioned in Bovens' framework (2007). The result is a hollowing out of political mechanisms intended to uphold ethical warfare and public legitimacy. In conclusion, the integration of AWS into the RNLA's operational architecture reveals a critical misalignment between emerging military technologies and traditional frameworks of accountability. Using Bovens' (2007) multi-dimensional model, it becomes clear that AWS dilute responsibility across a network of actors while simultaneously removing the central agent necessary for legal and moral reckoning. The resulting accountability gap is not merely a technical problem but a foundational

challenge to democratic governance and the rule of law. Without robust mechanisms for traceability, explainability, and human oversight, AWS risk institutionalizing impunity in modern warfare.

2.5 Meaningful Human Control

Meaningful Human Control (MHC) has emerged as a crucial concept to address the ethical concerns surrounding LAWS, particularly regarding accountability gaps, diminished human agency, and the potential for unintended harm (Verdiesen, Santoni de Sio, et al., 2021). First introduced by NGO Article 36 in 2013, MHC emphasizes that humans must retain ultimate responsibility over lethal decisions, ensuring compliance with ethical and legal norms (Amoroso & Tamburrini, 2021; Ekelhof, 2019). However, the precise definition and operationalization of MHC remain contentious, with scholars arguing that current interpretations often lack clarity and fail to ensure sufficient human oversight in practice (Blauth, 2023; Ekelhof, 2019; Kwik, 2022a). The next paragraph discusses several influential frameworks that try to operationalize MHC.

The operationalization of MHC proposed by Santoni de Sio and van den Hoven (2018) is grounded in a philosophical framework that seeks to ensure that moral and legal responsibility remains embedded within the design and use of AWS. Their model focuses on two core conditions, tracking and tracing, which together define the parameters under which human control can be deemed meaningful. Tracking refers to the capability of AWS to detect and respond to morally and legally relevant features in their environment, such as the presence of civilians, lawful combatants, and protected infrastructure. This dimension ensures that the AWS is not functioning in a morally vacuous space but is sensitive to contextual norms of warfare. Tracing, on the other hand, relates to the system's capacity to maintain an unbroken link between the actions it performs and the human agents responsible for its programming, deployment, and oversight. It is a demand for accountability infrastructure, ensuring that responsibility for any lethal decision can be attributed to identifiable human actors within the chain of command or system development process. Their approach is particularly significant because it moves beyond the simplistic dichotomy of human-in-the-loop versus human-on-the-loop paradigms. Instead, Santoni de Sio and van den Hoven propose that MHC be understood as a distributed ethical architecture. Rather than situating control at a single point, such as a trigger pull, they argue for a layered, systems-level view of control where ethical judgment and legal accountability are maintained throughout the lifecycle of the system. This aligns with broader concerns about the ethical risks of delegating lethal authority to non-moral agents. Their model insists that control must not be tokenistic but must involve informed, intentional, and morally aware engagement by human agents at key decision-making junctures. This conceptual model also finds support in the work of Christie et al. (2024), who emphasize the importance of traceability and explainability in ensuring that tracking and tracing are not only ethically justified but technically viable.

Christie et al. highlight that unless systems can generate decision logs, expose algorithmic reasoning, and facilitate post-hoc review, the tracing requirement becomes merely symbolic. This interdependence between ethical theory and technical design underscores the robustness of Santoni de Sio and van den Hoven's framework. One strength of this model lies in its flexibility. It can be adapted to various operational settings, technologies, and national doctrines while still upholding core ethical standards. However, critics might argue that it lacks concrete metrics or implementation protocols, potentially making it difficult to enforce in practical military contexts. Despite this, the model serves as a conceptual anchor in the MHC discourse, providing the foundational vocabulary and ethical grounding upon which other operational models, such as those by Umbrello (2021), or Cavalcante Siebert et al. (2023), build. By emphasizing moral sensitivity and legal attribution as preconditions for human control, Santoni de Sio and van den Hoven offer a compelling normative blueprint for regulating AWS. Their operationalization of MHC challenges both designers and policymakers to think holistically about the ethical landscape in which autonomous weapons will function.

Amoroso and Tamburrini (2021) propose a differentiated, principled, and prudential framework for MHC. Their model shifts the debate from the contested definitional ambiguities of autonomy toward the normative specification of human responsibilities over weapon systems. Central to their argument is the assertion that human involvement in AWS must consistently fulfil three critical functions: acting as a fail-safe actor, serving as an accountability attractor, and enacting moral agency. These roles ensure that humans maintain the ability to prevent unlawful harm, allow for the attribution of legal responsibility in cases of violations, and guarantee that life-and-death decisions remain traceable to human moral judgment, thereby safeguarding human dignity. Unlike approaches that advocate for a uniform standard of human control, Amoroso and Tamburrini argue for a differentiated model that adapts human control requirements according to the weapon's operational purpose ("what"), the context of deployment ("where"), and the technological capacities of the system ("how"). This differentiation is operationalized through a five-level taxonomy of human control, originally proposed by Sharkey (2016) and adapted by Amoroso and Tamburrini (2021), which organizes human involvement in AWS decision-making along a continuum. At Level 1 (L1), humans manually select and attack targets, exercising full control. At Level 2 (L2), machines suggest alternative targets, but human operators choose which to engage. At Level 3 (L3), machines autonomously select targets, yet humans must explicitly approve attacks before execution. Level 4 (L4) allows machines to select and engage targets under human supervision, where humans can intervene or abort if necessary. Finally, Level 5 (L5) denotes full machine autonomy, where the system selects and attacks targets without any human intervention post-activation. Amoroso and Tamburrini argue that higher levels of human control (L1 or L2) should be the default, with deviations permitted only in narrowly defined, internationally agreed-upon exceptions where structured environments and purely

anti-materiel targeting render lower levels (L3 or, in rare cases, L4) acceptable. L5 is categorically rejected as incompatible with the requirements of meaningful human control. Furthermore, their framework emphasizes the quality of human involvement, requiring the design of AWS to prioritize explainability and interpretability, and mandating comprehensive operator training to counter issues such as automation bias and erosion of accountability.

Building upon previously discussed tracking and tracing conditions, Cavalcante Siebert et al. (2023) operationalize meaningful human control through four actionable system properties. First, they propose defining a moral operational design domain (moral ODD), specifying not only technical limits but also ethical boundaries for system operations, ensuring alignment with societal norms and values (Cavalcante Siebert et al., 2023). For example, an automated vehicle might technically operate safely under certain conditions but should not do so if it risks harming vulnerable individuals. Second, the framework stresses the importance of establishing appropriate and mutually compatible representations between human and AI agents. Shared mental models concerning task distribution, environmental understanding, and mutual limitations enable the system to reliably track human moral reasons (Cavalcante Siebert et al., 2023). Third, the framework asserts that humans must possess both the ability and authority to influence the AI system's operation, ensuring they are empowered to act upon their moral responsibilities. Simply granting nominal authority without corresponding ability is insufficient; humans must be realistically capable of intervening when necessary (Cavalcante Siebert et al., 2023). Fourth, to satisfy the tracing condition, actions of AI agents must be explicitly linked to human decisions by way of explainable and inspectable connections. Cavalcante Siebert et al. (2023) emphasize the need for both forward and backward links: humans should be aware at the time of decision-making that their choices have moral consequences (forward link), and it should be possible to retrospectively trace any AI action to specific human decisions (backward link). Structured documentation practices, value hierarchies, and explainable AI techniques are proposed to facilitate these links and support accountability. Taken together, these four properties; moral ODD, shared representations, ability and authority alignment, and explicit human-AI linkages are deemed necessary (albeit not sufficient) for ensuring meaningful human control over AI systems. The framework thus shifts the discourse from abstract philosophical commitments toward tangible engineering and design requirements, advocating a socio-technical perspective that embraces transdisciplinary collaboration to address complex ethical challenges in AI deployment (Cavalcante Siebert et al., 2023).

Umbrello (2021) proposes a two-tiered framework that couples both the operational and design levels of abstraction to provide a holistic foundation for maintaining MHC. He argues that AWS cannot be ethically governed by focusing solely on operational control during deployment or solely on design intentions; rather, both aspects must be integrated under a systems thinking approach. The operational

level of abstraction emphasizes how military planning, mission briefings, rules of engagement (RoE), and target validation constrain the behaviour of AWS before and during missions (Umbrello, 2021). Here, the autonomy of AWS is not absolute but is embedded within pre-established military frameworks that shape and limit decision-making capacities. For example, pre-mission briefings detail objectives, target locations, collateral damage estimates, and weapon selections, creating layers of human input and oversight that frame AWS behaviour during engagements (Ekelhof, 2019). Thus, operational control demonstrates that "full autonomy," often feared in public discourse, does not translate into complete independence from human governance. Complementing this, the design level of abstraction addresses how AWS must be engineered to ethically align with human moral reasons. Here, Umbrello works with the tracking and tracing concepts created by Santoni de Sio & van de Hoven. Meeting these conditions ensures that AWS are explainable, transparent, and accountable even in complex operational environments. Systems theory and systems engineering serve as the overarching conceptual tools that connect both levels, viewing AWS not as isolated artifacts but as nodes embedded in larger socio-technical networks (Umbrello, 2021). This systems approach asserts that meaningful control emerges not from isolated human-in-the-loop interventions but from the ongoing co-construction of design values and operational constraints. Critically, Umbrello challenges the assumption that increasing autonomy inherently erodes MHC. If both operational and design levels are sufficiently robust, ensuring responsive behaviour and traceable accountability, then even technically "fully autonomous" AWS can remain under meaningful human control (Umbrello, 2021). This integrated perspective reveals that ethical concerns about AWS should not be focused narrowly on autonomy itself but on whether appropriate structures are in place across the lifecycle of development and deployment. Nevertheless, Umbrello acknowledges a limitation: the framework is better suited for pre-planned missions, such as aerial operations, rather than highly dynamic ground-based engagements where real-time adaptability might strain both operational planning and design responsiveness (Umbrello, 2021). Despite this limitation, the two-tiered framework significantly advances the debate by proposing a comprehensive, actionable model for understanding and achieving MHC. By coupling operational practices with ethically informed engineering, Umbrello (2021) offers a pathway to ethically integrating AWS into military operations without surrendering human moral authority.

While the above frameworks offer significant contributions to the conceptual and operational articulation of MHC, they often focus either on philosophical principles or practical task-level requirements without holistically addressing how responsibility can be continuously and dynamically maintained in a military operation. Ekelhof's (2019) analysis of meaningful human control in actual military operations, exemplified by the F-16 targeting case, challenges the notion that responsibility rests solely with the operator or can be guaranteed by design choices. Instead, her empirical work reveals that MHC is

fundamentally an organizational and collective responsibility, distributed across multiple actors, decision points, and institutional processes throughout the targeting cycle. Rather than seeing MHC as a single moment (e.g., the pressing of a button), Ekelhof demonstrates that crucial judgments, such as legality, proportionality, and target validation, are made at various stages, often well before the moment of engagement. This distributed approach underscores that human control is not merely a technical safeguard or procedural step, but a higher-order function embedded in the entire organizational architecture of military operations. This distributed, multi-actor conception of responsibility resonates closely with Bovens' (2007) framework, which conceptualizes accountability as a relationship between actors and forums encompassing information provision, debate, judgment, and possible sanctions, across political, legal, administrative, professional, and social domains. Both Ekelhof's findings and Bovens' theoretical model underline that genuine accountability, and by extension, meaningful human control, cannot be achieved through isolated individual interventions, but must be embedded within broader organizational and institutional structures. This recognition points to the limitations of frameworks that focus only on operator-level intervention or programming values. As autonomy increases and decision-making becomes further distributed across technical, organizational, and even political layers, meaningful human control must be understood as a shared, systemic function, requiring oversight, review, and input at each level of the targeting and deployment process.

It is precisely this broader, multi-layered view of responsibility that the CHOF brings into focus (Verdiesen, Santoni de Sio, et al., 2021). By explicitly including a governance layer, alongside the technical and socio-technical domains, CHOF provides an integrated structure for embedding ethical oversight, accountability, and collective responsibility across all phases of AWS development and use. CHOF thus operationalizes the insights of both Ekelhof (2019) and Bovens (2007): it moves from abstract calls for accountability to concrete institutional mechanisms, ensuring that meaningful human control is maintained not only at the operator and design levels, but throughout the organizational and governance layers of military action. True meaningful human control is not achieved through isolated interventions, but through robust institutional design, shared decision-making, and continuous oversight that permeate every level of military action, from the drafting of doctrine to real-time operational execution and post-mission review.

The CHOF is grounded in a dual-axis matrix that organizes oversight activities along both a temporal and structural dimension. Temporally, it considers the before, during, and after phases of AWS deployment. Structurally, it distinguishes between the technical layer (e.g., algorithmic parameters, system feedback loops), the socio-technical layer (e.g., operator interfaces, training protocols), and the governance layer (e.g., institutional norms, regulatory oversight). These dimensions are visualized in the 3×3 grid shown in

Figure 1 below, illustrating how comprehensive oversight requires actions and mechanisms across all intersections of time and structure.

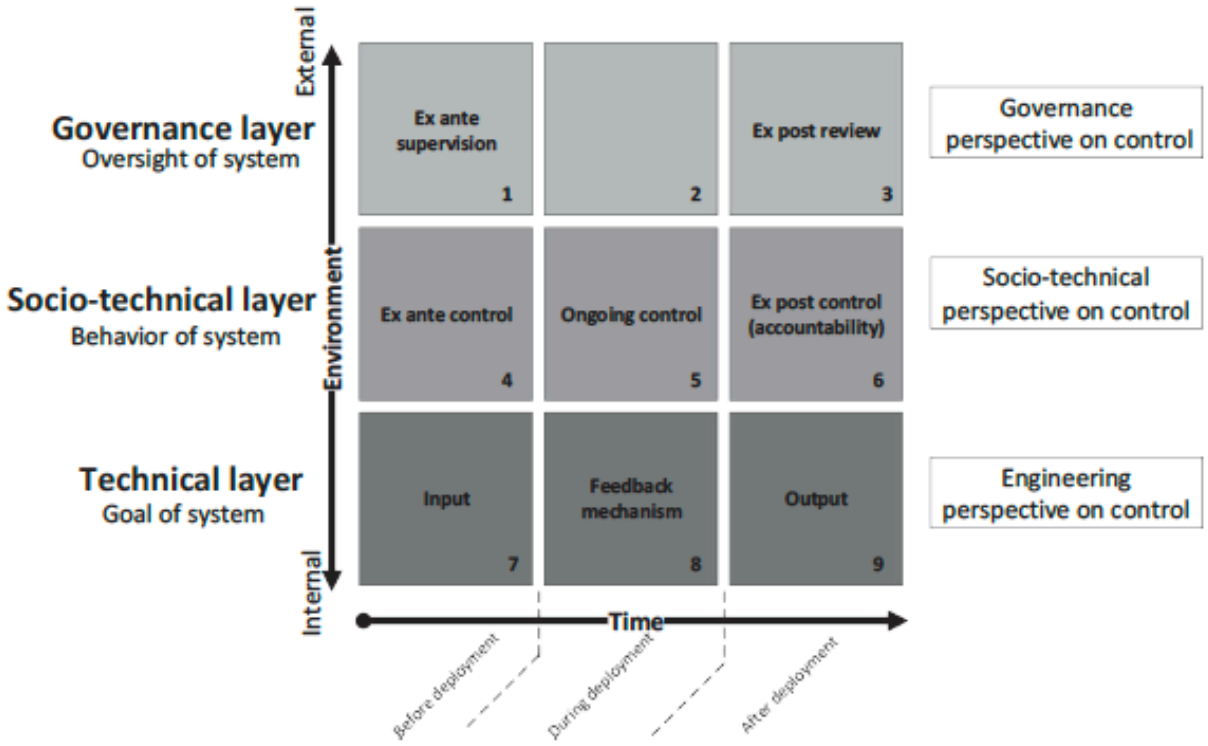

Figure 1. The Comprehensive Human Oversight Framework (CHOF) as developed by Verdiesen (2024), organizing oversight across time (before, during, after) and system environment (technical, socio-technical, governance)

Together, the nine blocks of the CHOF matrix function as an interdependent system: the governance layer sets mandates and rules of engagement, the socio-technical layer ensures that operators are trained, equipped, and empowered to act within those mandates, and the technical layer embeds the requirements into system design, testing, and operational performance. The temporal dimension reinforces this structure by requiring that oversight is not a single intervention but a continuous cycle of preparation, supervision, and evaluation. In this way, CHOF emphasizes that accountability cannot be meaningfully asserted if oversight mechanisms are absent at any intersection. Instead, the framework demonstrates how weaknesses in one block can undermine the entire oversight chain, while coordinated mechanisms across blocks can create resilience. Importantly, CHOF does not prescribe fixed solutions; rather, it functions as a diagnostic tool that reveals where oversight is strong, where vulnerabilities exist, and how institutional mechanisms may need to be reinforced. For instance, in the governance layer during deployment, the

moment of actual weapon operation, Verdiesen, Santoni de Sio, et al. (2021) identify a critical oversight gap: there is often no institutional mechanism actively monitoring or authorizing AWS behaviour in real time. This problem is compounded at the socio-technical level, where human operators may not have sufficient tools, situational awareness, or authority to intervene meaningfully.

To address the inherent opacity of AWS decision-making, particularly in systems powered by complex machine learning algorithms, Verdiesen, Aler Tubella, et al. (2021) proposes supplementing the CHOF with the Glassbox Framework. This integration is grounded in the recognition that the aspiration for full algorithmic transparency is, in most cases, technically infeasible. Advanced AWS often function as 'black boxes', with internal logic that is inaccessible to human auditors, whether due to technical complexity or proprietary constraints. The Glassbox Framework addresses this challenge by shifting the focus from internal system transparency to the generation and verification of observable, auditable behaviours through rigorous input-output monitoring. It is structured in two primary phases. The first, the Interpretation Stage, translates abstract ethical and legal values, such as proportionality, necessity, and respect for human dignity, into concrete, operational norms and requirements through stakeholder engagement. These norms are then embedded into system design and operational protocols, ensuring that critical values are directly linked to the system's expected outputs. The second, the Observation Stage, involves the continuous monitoring of AWS behaviour to assess conformity with these predefined criteria, enabling external, outcome-based audits even in the absence of insight into the system's internal logic. Notably, the Glassbox Framework does not reject the necessity of retaining certain black box elements; instead, it deliberately accepts this reality and ensures that accountability mechanisms remain accessible and meaningful to non-expert stakeholders, including military leaders, legal authorities, and policymakers. This pragmatic orientation allows for ethical performance to be externally assessed and audited, thus strengthening oversight. Structurally, the Glassbox Framework is integrated with CHOF: its Interpretation Stage aligns with CHOF's pre-deployment processes of formalizing values into design requirements, while its Observation Stage complements CHOF's deployment and operational oversight phases. Additionally, the outputs from Glassbox monitoring feed back into CHOF's review mechanisms, establishing a feedback loop that supports continuous learning and adaptive refinement of both technical systems and organizational processes. Together, these frameworks constitute an iterative, lifecycle-oriented approach to responsible and accountable AWS governance, bridging the gap between abstract ethical imperatives and practical, observable oversight (Verdiesen, Aler Tubella, et al., 2021)

.

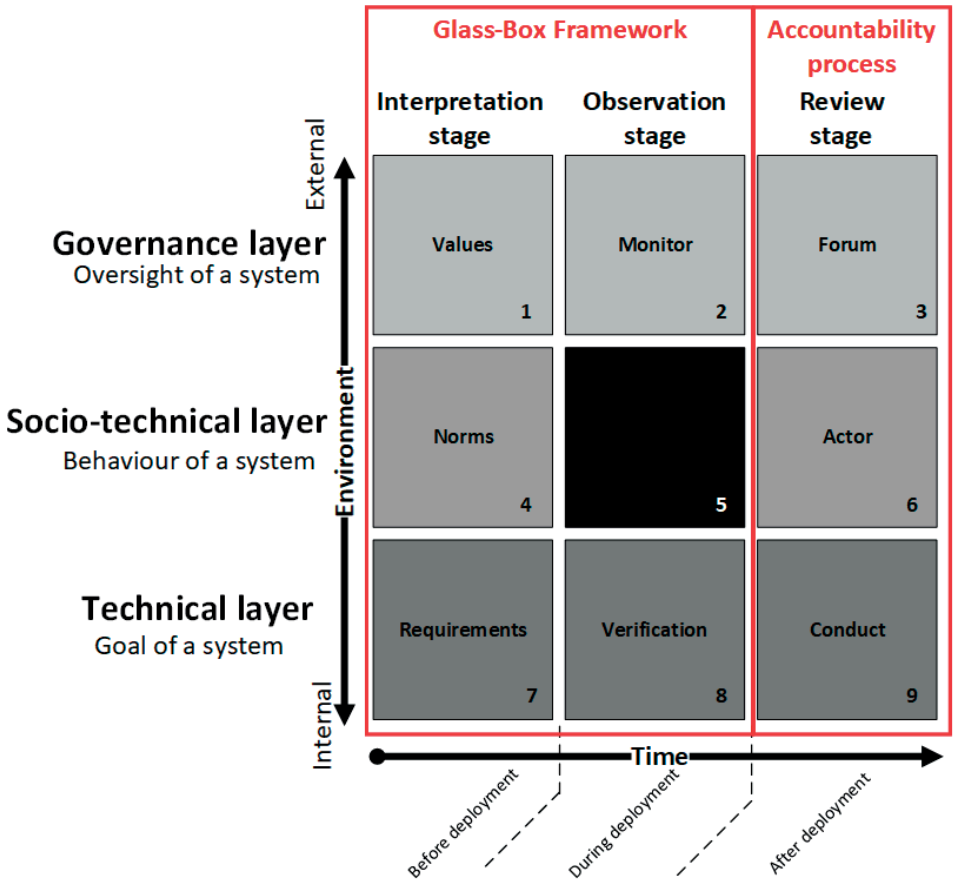

Figure 2: The CHOF with the Glassbox Framework projected on it

The CHOF offers a superior foundation for testing the ethical implementation of AWS within the RNLA compared to other available models. Unlike frameworks that centre only on operational or design phases, CHOF addresses ethical considerations across the entire lifecycle of AWS deployment by incorporating three interdependent layers: the technical, sociotechnical, and governance layers. This holistic approach aligns with the RNLA's emphasis on transparency, accountability, and compliance with both national and international legal standards (Defensie, 2023). As Verdiesen, Aler Tubella, et al. (2021); Verdiesen, Santoni de Sio, et al. (2021) argue, meaningful ethical oversight must extend beyond ensuring control during the moment of engagement; it must encompass system design, organizational practices, and institutional accountability mechanisms. In contrast, Amoroso and Tamburrini (2021) focus primarily on normative human roles during the targeting process, advocating for differentiated levels of human control depending on operational circumstances. While valuable, their focus does not address systemic issues such as organizational accountability and long-term governance structures, both of which are crucial for military institutions like the RNLA, which must operate under civilian democratic oversight and international humanitarian law. Similarly, the actionable properties proposed by Cavalcante Siebert et al.

(2023), although useful for engineering design, do not sufficiently address the broader institutional and sociopolitical dynamics that influence AWS deployment in real-world military contexts. Their focus on ensuring moral responsibility through design practices lacks the governance mechanisms necessary to ensure responsibility is meaningfully distributed across hierarchical structures, a need that is paramount in military organizations (Cavalcante Siebert et al., 2023). Umbrello's (2021) two-tiered approach, which integrates operational and design levels of abstraction, brings valuable insight into aligning design ethics with military operations; however, it still does not fully capture the governance requirements needed to safeguard accountability over the long term. In the context of the RNLA, which operates within a tightly regulated national and NATO framework, oversight must extend into political, legal, and institutional spheres, not just operational or engineering domains (Defensie, 2023, 2025; NATO, 2024). CHOF's governance layer, explicitly aimed at embedding oversight within institutional and societal structures, ensures that ethical concerns about AWS do not erode over time or become isolated within technical departments (Verdiesen, Santoni de Sio, et al., 2021). Moreover, the RNLA's commitment to responsible innovation and human rights norms necessitates a governance-driven approach that can be externally audited and internally validated, a capacity that CHOF directly supports through its multi-layered feedback mechanisms. In addition, Verdiesen, Aler Tubella, et al. (2021) emphasis on bridging gaps between value deliberation, human oversight, and technical development resonates strongly with the Netherlands' broader ethical AI initiatives (Defensie, 2023). Therefore, adopting CHOF provides a structured, value-sensitive, and practically implementable framework that ensures not only ethical design and deployment of AWS but also safeguards continuous accountability, public trust, and strategic legitimacy for the RNLA.

3 Methodology

This chapter outlines the methodological approach used to explore how the RNLA implements ethical oversight in the deployment of AWS, specifically through the lens of the CHOF. It describes the completed research paradigm, design, data collection and analysis strategies, and addresses ethical considerations and researcher positionality. The methodology was designed to bridge the gap between abstract ethical theory and operational military practice, providing a detailed and context-sensitive understanding of oversight mechanisms in AWS governance.

This study is grounded in an interpretivist research paradigm, which views reality as socially constructed and shaped by the subjective meanings that actors assign to their experiences (Alharahsheh & Pius, 2020). This paradigm is particularly appropriate for studying the ethical oversight of AWS, as it allows for an in-depth exploration of how actors within the RNLA interpret and implement abstract ethical principles in practice. The interpretivist approach supports a qualitative, case-based methodology, enabling the analysis of complex phenomena such as ethical accountability, institutional culture, and decision-making under uncertainty (Soeters et al., 2014). Interpretivism contrasts with positivist approaches by emphasizing context-specific understanding over universal generalizations (Alharahsheh & Pius, 2020). In this study, the focus is on understanding the practical application of the CHOF, rather than testing fixed hypotheses or measuring causal relationships statistically. This combination ensures that the research design and methodology reflect coherent ontological and epistemological assumptions, a key criterion for methodological validity in qualitative research (Soeters et al., 2014).

This research employed a single-case study design to examine how ethical oversight mechanisms for AWS are operationalized within the RNLA. A case study is well-suited for investigating context-bound, institutionally embedded phenomena where theoretical concepts, such as accountability, human control, and ethical governance, must be interpreted through specific practices and organizational cultures (Soeters et al., 2014). The RNLA provides a strategically relevant and theoretically rich case: it is an active participant in NATO, a proponent of technological innovation in defence, and an advocate for ethical leadership in military operations (Defensie, 2025). The case study approach facilitates a holistic and layered analysis of the implementation of the CHOF. It allows exploration of the manner in which abstract normative principles are translated into decisions, policies, and routines across CHOF's three layers: technical, sociotechnical, and governance. This level of granularity is essential for evaluating CHOF's real-world applicability and uncovering institutional conditions that either support or hinder its adoption. By focusing on the bounded system of the RNLA, the study adheres to the methodological standards of qualitative case research, providing both depth and internal validity. All data was collected

during April and May 2025, while the organization was adapting to AWS. While the findings may not be statistically generalizable, they offer analytical generalization by testing and refining theoretical constructs that can inform broader debates on military ethics and AI governance (Yin, 2009)

This study employed the CHOF as its principal theoretical lens. To ensure analytical rigor and consistency between theory and empirical inquiry, the CHOF was systematically operationalized into a coding and data collection structure. Each of the three layers of CHOF, technical, sociotechnical, and governance, was translated into concrete codes, thematic issues, and targeted interview questions (Verdiesen, Santoni de Sio, et al., 2021). For the technical layer, deductive codes included machine learning, traceability, and algorithm, capturing issues such as system design, data logs, and algorithmic transparency. Interview questions at this layer probed the extent to which machine learning processes were rendered traceable and how technical mechanisms either enabled or hindered oversight. The sociotechnical layer employed codes such as human control, operational use, and training, focusing on practical operator roles, the enactment of human control in practice, and training protocols for AWS deployment. Interview prompts explored personnel experiences with oversight and the adequacy of training and operational procedures. For the governance layer, codes included accountability, legal rules, and policy, mapping onto issues like Article 36 reviews, institutional accountability structures, and national or international policy development (Bovens, 2007; Verbruggen & Boulanin, 2017a, 2017b; Verdiesen, Aler Tubella, et al., 2021; Verdiesen, Santoni de Sio, et al., 2021). Interviews at this level addressed the institutional and legal frameworks governing AWS deployment and the mechanisms by which accountability is structured and implemented. A complete overview of interview questions can be found in appendix A. The initial coding scheme was deductively derived from the CHOF framework, with additional inductive codes incorporated during analysis to account for unanticipated themes emerging from the data (Braun & Clarke, 2006). This operationalization ensured that both document analysis and interview protocols systematically reflected the layered oversight model, allowing for a coherent and comprehensive thematic analysis of ethical oversight in the RNLA.

In this study, the CHOF served both analytical and organizational functions. It informed the structure of data collection instruments, such as the thematic focus of interview protocols, and has been a guide in the coding strategy used in qualitative data analysis. This layered approach allowed for a systematic examination of how ethical oversight is distributed and operationalized across the lifecycle of AWS in the RNLA. To enrich the analysis of the technical and sociotechnical dimensions, the study also incorporated the Glass Box framework as a complementary perspective. The integration of these frameworks allows for a more comprehensive evaluation of whether oversight mechanisms are not only normatively desirable but also technically feasible and verifiable. Together, the CHOF and the Glass Box framework

provide a coherent and multi-dimensional evaluative structure that supports both normative assessment and practical analysis. This dual-framework approach aligns with the interpretivist paradigm of the study by foregrounding how ethical oversight is not merely codified in policy but enacted through social practices, technical architectures, and institutional procedures.

This study draws on two primary forms of qualitative data: documentary analysis and semi-structured interviews. Together, these sources allow for a contextualized and interpretively rich exploration of how ethical oversight is operationalized in the deployment of AWS by the Dutch military. Triangulating institutional texts with expert perspectives enhances the credibility and depth of the research (Soeters et al., 2014). The documentary data comprised of 3 selected policy documents from the Dutch parliament, NATO and an advisory report from the AIV & CAVV. These texts provide insight into the normative, legal, and operational frameworks that guide ethical oversight of autonomous systems. Analysed thematically, the documents were mapped onto the three layers of the CHOF, technical, sociotechnical, and governance, allowing for a structured evaluation of how ethical principles are embedded across different levels of decision-making. Complementing the documentary analysis, semi-structured interviews are conducted with stakeholders selected through purposive sampling. This study employed a purposive sampling strategy to identify participants with direct relevance to the research objectives. Given the focus on the ethical governance of LAWS, individuals are selected based on their expertise and positional authority within domains central to the three layers of the CHOF. The goal is not statistical representativeness but analytical relevance, ensuring that participants possess the knowledge and experience necessary to illuminate the implementation of ethical oversight mechanisms within the Dutch military context. A total of eleven semi-structured interviews were conducted with participants purposefully selected for their diverse and often overlapping expertise in the development, deployment, and governance of AWS within and beyond the RNLA. Several participants held dual or hybrid roles that spanned operational command, technical development, legal advisory, and policymaking. For example, one interviewee was both a military captain and an AI modeler, explicitly bridging operational military practice with system development. Another combined legal expertise with operational duties at a drone battalion. The sample included policymakers from the Ministry of Defence (n=2), legal advisors (n=2), operational commanders involved in innovation or the implementation autonomy (n=4), military AI specialists (n=2), and an external expert (n=1), a Ukrainian drone commander whose insights were drawn from a related study. In several cases, participants' professional responsibilities spanned multiple domains; for instance, some commanders were also deeply engaged in technical innovation or policy formation. This deliberate inclusion of individuals with intersecting expertise allowed for a multi-layered exploration of ethical oversight as conceptualized by the CHOF framework. The sample size is guided by the principle of information power (Malterud et al., 2016), meaning that the number of participants is

adjusted in relation to the specificity of the research question, the theoretical framework employed, and the quality and depth of the data collected.

Data analysis in this study is grounded in thematic analysis, closely aligned with the interpretivist paradigm and the structural dimensions of the CHOF framework (Soeters et al., 2014). Thematic analysis facilitates the systematic identification and interpretation of patterns in ethical oversight practices as they emerge from the data (Braun & Clarke, 2006). This approach is applied to both documentary and interview data. Transcripts and documents are coded using a hybrid coding scheme: deductive codes are derived from the CHOF framework encompassing technical, sociotechnical, and governance layers while inductive codes are developed through close engagement with the empirical material. Coding is supported by qualitative data analysis software Atlas.ti, which enables systematic comparison of themes across data sources and participants (Friese, 2019). Attention is paid to both convergences and divergences in perspectives, revealing how ethical oversight is conceptualized, contested, and enacted within the Dutch military context. Through thematic analysis, the study goes beyond mere description by uncovering how ethical oversight practices are shaped within organizational settings, ensuring coherence between the study's theoretical commitments and empirical procedures (Braun & Clarke, 2006; Soeters et al., 2014).

In this study, validity and reliability were approached through the lens of trustworthiness, a concept widely adopted in qualitative research to ensure methodological rigor without relying on positivist assumptions (Nowell et al., 2017). This framework encompasses four key criteria: credibility, transferability, dependability, and confirmability (Lincoln & Guba, 1985; Nowell et al., 2017). These standards are particularly applicable in military research, where complex institutional dynamics and restricted access necessitate careful attention to methodological transparency and interpretive accountability (Soeters et al., 2014). Credibility was established through the triangulation of data sources and methods. By combining document analysis with semi-structured interviews, the study cross-validates findings and minimizes the risk of single-source bias. Member checks are incorporated during interviews, enabling participants to clarify or refine their responses, thereby improving interpretive accuracy. Transferability was enhanced by providing detailed contextual descriptions of the Dutch military's approach to autonomous systems and ethical oversight. Although the findings are case-specific, the conceptual insights may be applicable to other military institutions grappling with similar technological and normative challenges. Dependability was addressed by maintaining a comprehensive research log that documents the progression of analytical decisions, coding processes, and methodological adjustments. This audit trail supports procedural transparency and allows for external evaluation of research integrity (Creswell & Poth, 2016). Confirmability was ensured through systematic documentation of data interpretation procedures and the practice of reflexivity. As emphasized in military research contexts,

reflexivity involves ongoing awareness of how the researcher's position, particularly as a civilian studying a hierarchical and security-sensitive institution, may shape the framing of questions and interpretation of data (Soeters et al., 2014).

This study adheres to established ethical standards for qualitative research, with particular attention to the sensitivities inherent in military contexts and topics involving security, technology, and ethics. The research was designed to protect participants' autonomy, privacy, and well-being throughout all stages of data collection and analysis. Informed consent was obtained from all interview participants. Prior to each interview, participants received a clear explanation of the study's purpose, the topics to be discussed, their right to withdraw at any time, and the measures taken to ensure data confidentiality. Participation was entirely voluntary, and no coercion or incentives are involved. Consent was documented in accordance with institutional guidelines. Confidentiality was maintained through the anonymization of interview data and the secure storage of recordings and transcripts. Identifiable information is excluded from published findings, and pseudonyms or role-based descriptors are used where necessary to protect individual and institutional identities. All data is stored on encrypted devices and will be destroyed after the project concludes, in compliance with data protection protocols. The study also takes into account the potential for ethical dilemmas arising from the dual-use nature of military technologies and the classified status of certain information. Care was taken to avoid requesting or reporting on any information that could compromise national security or violate legal confidentiality. Participants were reminded not to disclose sensitive or classified content.

As the sole researcher conducting this study, I occupied a position external to the military institution under investigation. This outsider status affords a degree of analytical distance and enabled critical engagement with institutional practices without prior allegiance or embedded role expectations. However, it also introduced interpretive challenges related to access, insider knowledge, and organizational culture. The study acknowledges that all knowledge is situated, and that the questions posed, data collected, and themes emphasized are shaped by the researcher's background, interests, and normative commitments. In particular, my concern with issues of accountability, transparency, and human agency in technology governance influences the framing of research questions and interpretation of findings. These values align with the normative underpinnings of the CHOF but may also foreground certain aspects of oversight over others. Throughout the research process, I have adopted a stance of reflexive awareness, continually interrogating how my positionality affects interactions with participants, the interpretation of narratives, and the weighting of evidence. This included being attentive to potential power asymmetries in interviews with military personnel, as well as to the risks of projecting external ethical expectations onto a highly structured and operationally distinct institutional setting. The study embraced this reflexivity not as a

limitation, but as an epistemic strength. By explicitly acknowledging the interpretive lens through which the data is analysed, the research maintains transparency and supports the trustworthiness of its findings (Creswell & Poth, 2016; Soeters et al., 2014)

4 Results

The results of this study are structured according to the key dimensions of the CHOF (Verdiesen, Santoni de Sio, et al., 2021), which served as the central analytical lens for assessing how the RNLA manages ethical, legal, and operational oversight in the development and deployment of AWS. Drawing on qualitative interview data with policy makers, commanders, legal advisors, and technical experts, each theme explores how core principles of oversight, at the institutional, operational, and technical levels, are reflected in the RNLA's current procedures and organizational culture. In doing so, the findings illuminate both the established practices, and the emerging dilemmas associated with maintaining human control, accountability, and transparency in the context of evolving military technologies.

4.1 Theme 1: Institutional oversight

A central element of human oversight, as articulated in the CHOF, is the embedding of robust ethical and legal norms within the doctrine and professional identity of the armed forces (Verdiesen, Santoni de Sio, et al., 2021). Within the RNLA, the foundational principles of IHL, including distinction, proportionality, and necessity, are not only codified in policy documents but are also integral to the Army's operational routines and sense of professional responsibility. This is reflected across strategic, operational, and individual levels. Both policy makers and operational commanders emphasized that the development and deployment of AWS are consistently guided by the core tenets of IHL. As one policy maker explained: *"You initially translate the basic principles of humanitarian law of war. That's actually the basis, so you have the principle of distinction, principle of proportionality. Yeah. So first of all, you have to know, as a commander, can such a system, what does it take to be able to judge that such a system is deployed in accordance with International Humanitarian Law, and that's how you're going to reason back in making policy."* This orientation is echoed by commanders involved in operationalizing autonomous concepts, who noted: *"The obligation must be sufficient discrimination on a legitimate target. Along with adherence to International Humanitarian Law. No one questions that. That is not only an ordered obligation, but also a shared feeling within us soldiers. That's how I want to work. It's part of the identity of being a professional soldier. Yes, that's how I have experienced that myself. So, we also express this feeling as soldiers."* This professional internalization of legal and ethical norms was described not simply as compliance, but as a foundational aspect of military identity and purpose. One respondent reflected: *"Like Clausewitz said, war is the continuation of politics. And how you conduct war will support the political agenda. And so, the deployment of fully autonomous systems may undermine our long-term political agenda. And that is something that has to be weighed against being forced to undermine our political agenda by our opponent. Because our opponent is also doing a political agenda and trying to force us to*

let go of ours. So, if we fall into their trap of engaging in warfare the way they do, then we lose what we were fighting for, and we might as well not fight. So that's also, I mean, that applies to abiding by IHL, all of it. Because if you don't, IHL in and of itself puts us on a weaker standing. Especially against an opponent like Russia, who clearly has no respect for IHL. But if we don't stand by it, then why are we even fighting? Because then we're basically them. So that's the broader thing that people need to remember when we discuss the ethics of what we're doing."

These findings illustrate the extent to which ethical and legal norms, framed by both IHL and the CHOF's emphasis on institutional and operational oversight, are not only formal requirements but also serve as guiding values shaping the Army's collective and individual conduct. In this way, the foundational integration of IHL forms the bedrock for all subsequent procedures and accountability mechanisms, providing a clear point of reference for both policy and practice as the Army adapts to new technological realities. The interviews and policy documents indicate that the RNLA employs a series of institutional procedures to ensure compliance with IHL throughout the lifecycle of weapon systems. These procedures are typically structured around three phases: pre-use (anticipatory review), use (operational deployment), and post-use (evaluation and accountability).

Pre-Use Oversight

A foundational mechanism for ensuring adherence to IHL within the RNLA is the anticipatory legal review of new weapons and technologies, mandated under Article 36 of Additional Protocol I. This review serves a critical preventive function, aiming to identify and address potential legal or ethical concerns before a weapon system is ever fielded. As one interviewee underlines, "*It is an obligation that new means and methods of warfare must be tested for compatibility with the laws of war, during study, design, and development."* This requirement anchors legal and ethical scrutiny as an integral part of the military's development process and establishes Article 36 as a central safeguard in the governance of military innovation. Responsibility for conducting these reviews lies unequivocally with the Ministry of Defence. Even when weapon systems are developed externally, the Ministry retains final authority over legal compliance. This responsibility is distributed across several departments, including the materiel organization and legal advisors, who are collectively tasked with testing, verification, and final approval. As one technical manager explains, *"Responsibility for whether the neural network performs well lies with test, evaluation, verification, and validation, organized at the materiel organization. Not different from conventional weapons."* In practice, the final legal review is reserved for Defence jurists, while technical and operational input may be sourced from other experts, such as those at TNO during early research phases.

However, the introduction of AWS has complicated the practicalities of Article 36 reviews. Legal experts and practitioners highlight significant challenges around transparency and explainability, features that are critical for robust legal assessment but can be difficult to guarantee with complex or proprietary AI systems (Christie et al, 2024). One legal expert expresses this ongoing dilemma: *"If you start building autonomous weapon systems that can independently select and engage targets, it is still an open question as to which parts of the software need to be transparent to your organization in order to conduct a legal review. That obligation comes from Article 36... These are all issues for which, honestly, much still needs to be resolved."* Another interviewee further stresses, *"If, in our opinion, they cannot clearly demonstrate explainability or traceability, we will not buy or use it."* This insistence on transparency is not just a matter of legal formality; it reflects deeper ethical and operational concerns about accountability in machine-driven targeting. What complicates the Article 36 review is depends not only on the system itself but also on the intended usage scenarios. As one expert from TNO notes, *"A very important part of such a weapon review is to define the usage conditions, when you can use it, under what circumstances, and how. This applies to all weapons systems."* The context in which an AWS will be deployed shapes the requirements for legal and ethical oversight, making it essential that Article 36 reviews are tailored to specific operational realities and not treated as generic procedures. However, the interviews show that it is yet unclear how these usage cases will be shaped, and that complicates the process of the article 36 review. Several respondents acknowledge that procedures are still under development as Defence seeks to formalize policy and translate legal principles into concrete operational guidance: *"That policy vision will eventually be developed further, so that the people within Defence who actually work with autonomous weapon systems can consult it... It will include the principles we want and specify what the system must meet legally and ethically,"*.

In summary, Article 36 reviews constitute a foundational safeguard for ensuring IHL compliance in the development of new military technologies. However, with the increasing complexity and autonomy of weapon systems, new demands are emerging around transparency, explainability, and context-specific oversight, challenges that are not yet fully resolved in practice. Importantly, many of these issues are closely tied to the actual operational scenarios in which autonomous systems are used. For this reason, a more detailed analysis of Article 36 reviews in relation to specific usage cases will follow in a later section of the results. For now, it is essential to recognize that Article 36 reviews form only the first layer of legal and ethical oversight. The next critical procedural safeguard is the targeting cycle, which governs the decision-making processes during the operational deployment of weapon systems. The following section examines how the targeting cycle is structured to uphold legal and ethical standards in practice, and how it interacts with the unique characteristics of autonomous weapon systems.

Operational Oversight

The targeting cycle stands as the primary institutional mechanism through which the RNLA ensures that the use of force remains compliant with IHL. The targeting cycle is not a mere formality but is deeply institutionalized in Dutch military practice. Its structured sequence is explicitly designed to guarantee that all uses of force, including those involving AWS, are subjected to rigorous legal, ethical, and operational scrutiny. As one legal advisor notes: *"The targeting cycle, very simply put, first starts with identification of the targets. Then, when we have identified the targets, we will select which targets we want to engage. Of course, you will then have to decide how to engage the target. This gets evaluated very stringently, all the way up till your weapon choice, the angle of entry, does it have a delay fuse, etcetera. In this process you make a collateral damage assessment, so you estimate how much collateral damage you expect for each choice of weapon. Then, of course, you will select the weapons with the most effectiveness for the least amount of collateral damage. This then also entails the legal check: can we justify the collateral damage in line with the military benefit we receive? Yes? Then we press the button and the attack is commenced. This then is evaluated: did we manage to do what we anticipated?"* By structuring each stage of target selection, engagement, and post-action review, the targeting cycle operationalizes the core IHL principles of distinction, proportionality, and necessity, not just as abstract requirements, but as real-time decision criteria that must be satisfied before any use of force is authorized. This procedure is reinforced by extensive training and legal support for operational personnel, ensuring that all actors are equipped to apply these standards in practice. As another respondent emphasizes: *"Those are actually the regular ways we already have, before we use weapons, to make sure it happens according to the laws of war. So, it's embedded in the targeting cycle. There is training for the operational people, there is legal support involved."* Moreover, the targeting cycle is deliberately designed to be adaptable to varying operational contexts. In cases of deliberate targeting, such as strategic strikes with extended planning horizons, the process allows for comprehensive intelligence gathering and legal review. As a policy expert notes: *You can take a long time to do that. Sometimes weeks, or even months. Then you have a legal advisor that gives judgment on the targeting,"* Conversely, in situations requiring dynamic targeting, such as sudden ambushes, decisions must be made rapidly and often with incomplete information: *"Dynamic targeting happens much faster, for example when soldiers are suddenly ambushed, and you have way less information to make the legal decision. As a result, the legal decision, whether it is proportional or not, has to be made at a lower level. That is a completely different situation."*

Given that the targeting cycle fundamentally relies on human moral reasoning, contextual awareness, and experiential judgment, the implementation of autonomous weapon systems within this process raises pressing concerns. Integrating AWS into the targeting cycle means introducing systems that are inherently

unable to mirror the depth of human ethical deliberation or adapt to novel and ambiguous circumstances. This disconnect threatens to undermine the targeting cycle's capacity to uphold the principles of IHL in practice. One officer explained, *"Those are the rules of the laws of war that give substance to this. If I want to attack a target, then it must comply with, broadly, the principles set by the laws of war. I make sufficient distinction between civilian and military. I am allowed to hit soldiers, not civilians. There must be military necessity. If I want to kill my enemy, that is allowed. But if I can arrest him, it remains a human being. Arresting is also an option."* Such moral reasoning, where a commander weighs not just the rules but the context and consequences, cannot be reduced to algorithms: *"The degree of foreseeable collateral damage, whether or not you want to accept it, is a human judgment. It depends on the target you want to hit and on the surrounding circumstances, time, experience, and so on. If I can eliminate a very important target, then I am willing to accept more collateral damage. If that target is not so significant, I am basically willing to accept less damage. A device cannot assess that. It simply cannot."* The inability of AWS to replicate this depth of ethical reflection is particularly acute when acceptable harm in technical terms would, for a human, remain impermissible. As another respondent observed: *"You can keep thinking about what is and is not humane. Causing unnecessary suffering in general is a no-go. 'Unnecessary' is a word you can philosophize about for a long time. What is not necessary? That's a judgment you make based on years of experience, humanity, science, being a good person. But when is something unnecessary? If you ask an AI system, you are immediately lost. It does not work. So, these kinds of considerations, what is unnecessary suffering, what is the right distinction, what is acceptable collateral damage, what is not, all these are the product of human thought."* Empathy and contextual reasoning are not the only qualities that AWS lack; the technical limitations are equally significant. As a military programmer highlighted: "*All those beautiful AI systems, they are non-causal systems. They are good at finding correlations, but they are not good at reasoning in terms of cause and effect. Especially in military, creative scenarios… you need a human with intuition."* Humans can interpret ambiguous, novel, or unexpected situations such as determining whether a concealed object signals a threat or civilian presence while machines remain restricted to their programmed experience: *"Only a well-trained person, who has experienced such a scenario before, can assess that correctly. And even then: in the moment, you still have to recognize it."* Further, *"You also cannot test for things you do not expect. We expect aluminium foil now, so you can test for that. But if you later encounter a completely new type of tank, even the number of road wheels is unknown. How is a system supposed to recognize that?"* This technical limitation extends far beyond the battlefield. One programmer makes a simple example saying that an autonomous system would not even understand the context of doing grocery shopping: *"Even just going grocery shopping, how much reasoning is involved! A robot must be told*

everything. What is bread? Why do you want it? Which supermarket do you choose? How much money do you have? That kind of practical knowledge is extremely difficult to transfer to machines."

In sum, implementing AWS in the targeting cycle risks eroding the very safeguards that are essential for IHL compliance, precisely because these safeguards are built upon qualities of moral discernment, situational awareness, and intuition that remain beyond the reach of current autonomous systems. It is important to emphasize that the implications of integrating AWS into the targeting cycle are not monolithic; rather, they are shaped by a combination of contextual factors and design choices. The operational environment, whether conventional or asymmetric, stable or fluid, affects the risks associated with delegating specific targeting functions to autonomous systems. Equally crucial is the level and placement of autonomy within the targeting cycle: for example, assigning AWS responsibility for preliminary target identification poses fundamentally different challenges than permitting fully autonomous weapon release without human intervention. Therefore, the risks to IHL compliance are highly variable and must be assessed not only by examining the technology itself but also by considering where and how autonomy is implemented within military decision-making processes. This is further explained in theme 2.

The following section examines how these institutional challenges concerning the targeting cycle extend into the post-engagement phase, focusing on accountability mechanisms such as After-Action Reviews and the evolving requirements for transparency and traceability.

Post-use Oversight

Accountability is a cornerstone of ethical and lawful military conduct within the RNLA. It serves not only as a check on individual and organizational behaviour but also as a safeguard against unlawful harm and excesses in the use of force. As a military lawyer specializing in drone warfare explains: "*Every action you perform must be accounted for. Without accountability, a governmental organization would be unable to amount to something. […] If you let soldiers operate without some form of accountability, you can expect that people will get unjustifiably hurt. You can see it in Ukraine now, with that attack on Sumi. Look, these kinds of acts, every human can do that. We are not morally superior; every man can do that. But the question is, do you want that as a society? Well, we do not what that. Maybe in Russia they want it. But here, we wouldn't. And that accountability, will push people to keep on thinking, even when times get bad. Every carpenter can have a bad hit. Every soldier can take a shot he shouldn't have taken. That is okay, everyone makes mistakes. But when it gets excessive, it must be called to a halt."*

The RNLA's accountability process begins with two foundational procedures: the After-Action Report (AAR) and, where applicable, the Battle Damage Assessment (BDA). These instruments serve as the

primary means by which operational decisions and their consequences are documented, forming the first link in a broader chain of accountability. The AAR is a procedural requirement following every engagement, regardless of the outcome. A commander is tasked with compiling this report, based on immediate observations and initial documentation such as the Troops in Contact Report. As one interviewee describes: *"He will do that based on his Troops in Contact Report. There, in some keywords, the commander writes down what has happened. Then in the After-Action report, the commander writes this operation had these results, this done. Some reports get written down very well, others do not."* This illustrates both the standardized intent of the procedure and the variability in reporting quality.

Importantly, AARs are always required, while BDAs are situational. As one participant notes: "*You always have an after-action report. You don't always have a battle-damage assessment."* Whereas AARs capture all available information at the moment of engagement, acknowledging operational uncertainty: "*We could see through our scopes that we shot two people... but we're not sure."* the BDA serves a different, more investigative function. The BDA is carried out after the fact, aiming to determine the actual effects and outcomes of weapons use. This may involve returning physically to the site of engagement or leveraging remote sensing technologies such as drone or satellite imagery. BDAs can be conducted hours or even days after the event, allowing for a more objective evaluation of damage and casualties. The BDA may reveal discrepancies with the initial AAR, as described in one interview: *"They didn't appear to be Taliban fighters after all, but civilians."* Such findings can trigger further review and potentially corrective actions. In some cases, the BDA consists solely of visual evidence: *"A battle damage assessment is sometimes just the image data."* This was significant in high-profile cases such as Hawija, where the temporary loss and recovery of BDA imagery affected the transparency of the accountability process.

While the AAR and BDA form the essential starting point of the Royal Dutch Army's internal accountability process, they are only the first link in a comprehensive, multi-layered chain of oversight. To ensure that operational actions are subject to independent scrutiny and legal review, the accountability process extends well beyond the unit and command structure. After every operation involving the use of force, the procedures for transparency and traceability are reinforced through the active involvement of the Koninklijke Marechaussee (KMAR) and the Public Prosecution Service (Openbaar Ministerie), creating a robust interface between military operations and civilian judicial authority. As one respondent explains: *"An after-action report is sent to The Hague and also to the Public Prosecutor via the KMAR officer."* The KMAR officer embedded within missions is responsible for forwarding all reports of weapons use to the Public Prosecution Service, which can then decide whether further legal investigation is warranted. If the Public Prosecution Service suspects a violation of IHL or criminal conduct, it may

initiate a formal criminal investigation, at which point the Ministry of Defence is required to halt its own internal inquiry. As one interviewee describes: *"If it appears that we are dealing with criminal conduct, Defence must stop its own investigation.* "This strict separation of powers is designed to safeguard the independence and impartiality of legal proceedings, ensuring that allegations of unlawful conduct are not solely handled within the military hierarchy. The dual reporting system reflects the hybrid civil-military nature of the Dutch accountability structure, combining operational oversight with legal traceability. This process aligns with Bovens' (2007) multi-dimensional model of accountability, which emphasizes the need for political, legal, and administrative mechanisms. In addition to standard reporting and legal review, the Dutch Ministry of Defence maintains its own internal mechanisms for investigating the use of force. These internal inquiries may be initiated independently of, or in parallel with, legal investigations by the Public Prosecution Service, depending on the nature and severity of the incident. As one respondent notes: "*Within the Ministry of Defence it is possible to conduct further research, that is an internal weapons usage investigation"* Findings from these investigations are structured along a scale of legal and operational responsibility, ranging from complete procedural compliance to the most serious breaches of international law. The gradation reflects the RNLA's need to assess both the legality and the ethical quality of actions taken during military operations. Cases where no civilian casualties occurred and all procedures were correctly followed require no further action beyond documentation: "*The first possibility is: we investigate, and we cannot establish that civilian casualties occurred. So that's it."* The second possibility arises if civilian harm is found but all procedures were followed and a legitimate military target was present which means the operation is deemed lawful: "*We see that civilian casualties occurred, but all procedures were followed correctly. So, the attack was lawful*" In such cases, the Ministry may choose to offer ex gratia compensation to victims or rebuild destroyed property: "*We can choose to compensate the surviving family members voluntarily... but that's out of our own will, not because we are obligated to."* If further investigation reveals the attack was unlawful, due to misidentification of targets or intelligence failures, responsibility is escalated, even if there was no criminal intent: *"We attacked this house... with the information we have now, we know civilians were there. So, we should never have attacked it*" Such cases may result in administrative or policy changes, and if criminal conduct is suspected, the Public Prosecution Service takes over the investigation. Deliberate attacks on civilians or disproportionate use of force are treated as war crimes, demanding immediate transfer of authority to civilian prosecutors: *That's when, for example, a school is deliberately targeted... or one soldier is killed, but a hundred civilians die. That is not proportionate.* As clarified by a respondent: "*If that's the case... then Defence stops its investigation. The Public Prosecution Service takes the lead."* This structured, tiered approach to internal accountability not only reinforces compliance with IHL but also enables the Ministry to respond proportionally to the severity of incidents, providing a basis

for learning, redress, and, where necessary, reform. This gradation of decision-making and investigation supports Bovens' (2007) conception of accountability as a process that goes beyond formal rules, requiring both judgment and proportional response. It also illustrates the need for traceability and explainability, key principles in the literature on meaningful human control and the CHOF framework (Verdiesen, Santoni de Sio, et al., 2021), especially as complexity increases with the introduction of autonomous systems.

While the formal procedures for accountability in the RNLA are extensive and multi-layered, the empirical data reveal that their effectiveness is frequently shaped, and at times undermined, by the underlying reporting culture and organizational dynamics. Despite the presence of mechanisms such as AARs, BDAs, and legal reviews, the willingness of personnel to report fully and transparently is influenced by a complex interplay of loyalty, solidarity, and perceived risk. Interviewees consistently highlighted that soldiers often experience a tension between institutional duty and peer solidarity, which can result in selective or incomplete reporting. As one participant observed: "*Because soldiers may feel that reporting truthfully means turning in a comrade."* This tension is particularly acute in situations where admitting mistakes or uncertainty could lead to disciplinary consequences or damage the reputation of the unit. The Hawija case provides a salient example: although civilian casualties later emerged as a central issue, initial reports marked "none" for collateral damage, rather than the more accurate "unknown." As the interviewee remarked: *At Hawija, 'none' was recorded rather than 'unknown', and that raises the question: why?* This suggests not merely administrative oversight, but potential strategic framing of information to mitigate scrutiny. Recognizing these vulnerabilities, the Dutch Ministry of Defence has initiated reforms aimed at strengthening the culture of reporting and ethical reflection. Recent changes include the incorporation of moral and operational dilemma scenarios into training exercises, thereby normalizing ethical discussion and transparency before deployment: *"We aim to include these dilemmas in training exercises so soldiers can talk through them."* Moreover, additional channels for anonymous reporting have been established, such as confidential advisors and external organizations like the Veterans Institute, to provide personnel with safe avenues for raising concerns outside the chain of command. Nevertheless, these reforms are not without challenges. As one respondent noted: *"That is a major question, how do you ensure people fill in such reports correctly?"* This underscores a fundamental insight from the literature: the efficacy of accountability mechanisms is not determined solely by their formal design, but by the internalized norms, integrity, and trust of the individuals responsible for their execution (Bovens, 2007; Verdiesen, Santoni de Sio, et al., 2021). Where a culture of transparency and critical self-reflection is lacking, even the most well-structured accountability systems may prove inadequate.

This accountability chain has been thoughtfully created to ensure accountability within the Netherlands Army. However, the introduction of autonomous weapon systems fundamentally challenges the very foundation of this accountability chain. Traditional military accountability is built on the principle that responsibility for decisions and actions can be traced back to individual persons; those who plan, authorize, and execute operations. The entire system of AARs, BDAs, legal reviews, and subsequent disciplinary or legal consequences is designed to attribute blame or exoneration to specific persons. Yet, as autonomy in weapon systems increases, the direct human link in decision-making becomes more ambiguous or even absent. This creates significant friction with existing accountability mechanisms, which presuppose the presence of an identifiable actor who can be questioned, sanctioned, or held legally liable (Bovens, 2007). As autonomous systems have the ability to take on greater roles in critical functions such as target identification, selection, and engagement, the established frameworks for attributing responsibility, and therefore for enforcing justice, are increasingly strained, if not rendered obsolete (Ebrahimi, 2024; Umbrello, 2021; Wood, 2023). This shift exposes a fundamental tension: while accountability mechanisms remain vital for upholding legal and ethical standards, they were not designed to accommodate the diffusion or displacement of agency that occurs with advanced automation and artificial intelligence.

This diffusion of agency can provide to be a problem for the ethical implementation of such weapon systems, as one interviewee succinctly notes: *"Part of the ethical implementation is accountability. So that's the key part of being able to ethically implement something, pass the laws of discrimination under IHL, but also accountability. The law is only as strong as your ability to enforce it. And we don't have a way of enforcing accountability on autonomous systems yet."* This resonates with Boven's (2007) ideas of accountability as a form of responsibility. It is even more important if looked at the fact that, as seen in the previous section, there is already a lot of doubt whether AWS can reliably make life or death decisions. Whenever there is doubt or complexity in the targeting cycle, the need for a functioning accountability system is even more amplified. The accountability mechanisms are designed to ensure that responsibility for military action can always be traced to identifiable human actors. This not only allows for punishment in cases of unlawful conduct, but also for the broader societal goal of ensuring trust in the use of force (Bovens, 2007). The interviews show that these traditional forms of justice rely on the ability to identify, punish, or deter responsible individuals: *"People are held responsible. You can punish companies, with a fine, for instance, but a company cannot feel pain. A company can be insulted. A company cannot… feel sad. A company cannot physically feel pain. Nor can a system. Nor can AI. And the whole idea we have of accountability: revenge. If someone does something wrong on purpose… that person must suffer for it. That is the entire essence of criminal law: retribution. But that is not possible.*

With an autonomous system, you cannot exact retribution. And to me, that's a sticking point. That's a problem."

Thus, the integration of autonomy into the human-centred accountability cycle presents a fundamental and immediate threat to the very basis of military accountability. From a policy perspective, current international legal frameworks maintain that states remain accountable for the actions of their autonomous weapon systems. Dutch policy, for example, emphasizes the importance of clear responsibility in cases of unlawful use, stating: "*In the development and deployment of partially autonomous weapon systems, it is of great importance that it is clear where responsibility lies in the event of unlawful use. Under general international law, states can be held liable for the unlawful actions of autonomous systems they employ, for example when these systems target civilians. Under international criminal law, individuals can also be held liable who play a role in the use and development throughout the lifecycle of a partially autonomous system, particularly developers, commanders, and operators."* (AIV&CAVV, 2021). However, simply assigning accountability to the state is an insufficient solution, both practically and ethically. As one respondent observed, state responsibility may offer a legal avenue for reparations, but fails to address the need for meaningful justice or individual culpability: *"The degree of accountability you cannot impose with autonomous systems means you are actually reconfiguring the whole accountability system... If you do have an autonomous system, say the Patriot shoots down not the enemy but MH17... the only thing you can do is convict the state that used it to reparations. More than that, you actually have nothing. And if you think about it that way... then all the harm caused is mechanized. The one responsible for programming... also goes free."* More troubling still, the diffusion of agency across the development and operational lifecycle of AWS makes it nearly impossible to assign accountability to any single individual. The traditional model of criminal justice, which relies on the identification and sanctioning of specific human actors, is fundamentally undermined by the collaborative, multi-layered nature of autonomous system deployment. As another interviewee explained: *"There's the classic discussion, is it the programmers? Is it the company as a whole? Is it the commander who deployed it that is responsible? And also, which company, because these things are normally fully autonomous systems, may have multiple software's on it from different companies. So, is it one specific piece of that autonomous system that went wrong? Is it the system as a whole that went wrong?"* In a typical military context, responsibility might be assigned to commanders or states. But in a multi-layered, globalized technological ecosystem, where software and hardware come from different providers and may interact in unpredictable ways, tracing a causal chain from action to actor becomes exceedingly difficult.

The analysis of the RNLA's procedures reveals a robust, multi-layered architecture for ethical and legal oversight, organized around three distinct phases: pre-deployment (anticipatory review and Article 36

assessments), operational deployment (the targeting cycle), and post-deployment (accountability processes such as AARs and BDAs). Each of these mechanisms was developed on the premise of continuous human involvement and control, ensuring that decisions about the use of force are embedded in professional judgment, contextual awareness, and institutional values rooted in IHL. However, the empirical findings demonstrate that the introduction of autonomous weapon systems fundamentally jeopardizes each layer of this oversight structure. In the pre-deployment phase, the opacity and complexity of AWS challenge the effectiveness of Article 36 reviews, making it increasingly difficult to ensure transparency, explainability, and context-specific legal compliance. During deployment, the targeting cycle**,** a core mechanism for operationalizing IHL principles**,** becomes vulnerable when moral reasoning, empathy, and situational judgment are replaced or diluted by automated decision-making. Post-deployment, the diffusion of agency and ambiguity in responsibility associated with AWS threaten the integrity of accountability processes, undermining both individual and collective liability, as well as public trust in the ethical governance of military force. What ties these vulnerabilities together is the progressive erosion of meaningful human involvement. The effectiveness of oversight mechanisms is predicated not just on their formal existence, but on the sustained presence of human judgment at each decision point. The introduction of autonomy, when not properly constrained or supervised, disrupts this dynamic, creating significant gaps in ethical, legal, and operational control. As a result, the very foundations of military accountability and adherence to IHL are called into question. This analysis highlights the urgent need to reexamine and reinforce the principle of human control across all phases of AWS development and deployment. The concept of human control**,** its definition, scope, and practical implementation**,** will be explored in depth in the next theme, focusing on how autonomous weapons can and should be used in military operations, and what forms of human oversight are necessary to safeguard both legal and ethical standards in the evolving landscape of armed conflict.

4.2 Theme 2: Operational Usage

The conceptualization of AWS within the RNLA is rooted in both international legal norms and the evolving demands of military operations. Dutch policy, closely aligned with influential academic and humanitarian literature, generally defines AWS as "*systems that after activation, possess the ability to select and attack targets without further human intervention"* (AIV & CAVV, 2021). This definition is echoed by the majority of interviewees, who describe AWS as: "*a system that, once activated, can independently select and engage targets."*

A critical distinction is made between unmanned and autonomous systems. Unmanned systems, such as remotely piloted drones or ground robots, operate at a distance but remain entirely under human control for navigation and engagement (Altmann & Sauer, 2017; Verbruggen & Boulanin, 2017a). By contrast,

autonomous systems are designed to execute specific tasks, such as navigation, target identification, or engagement, without direct human input after activation, relying instead on pre-programmed criteria or adaptive algorithms (ICRC, 2019). This difference is not merely technical but shapes the operational, legal, and ethical contours of their use. As one technical officer charged with innovation within the army put it: *"What I see as autonomy is that you give a task to a system, and within that task, it receives a certain degree of decision-making, which you as a human determine in advance. That is crucial: the human determines how much decision-making freedom the system gets."* The literature and policy discourse further clarify this spectrum of autonomy by the degree of human involvement, using terms such as "man-in-the-loop" for direct control, "on-the-loop" for systems under human supervision with the possibility of intervention, and "out-of-the-loop" for systems that function fully autonomously without human input during action (AIV & CAVV, 2021; Christie et al., 2024; Longpre et al., 2022). These gradations are central not only for technical and tactical planning but also for the ongoing debates about meaningful human control and the limits of responsible delegation. Dutch doctrine and policy emphasize that technological capacity for autonomy must be matched with robust human oversight and accountability. The prevailing view, reflected both in interviews and the policy documents, is that autonomy in weapon systems must never come at the expense of the ability to understand, direct, and, where necessary, override machine decisions (Amoroso & Tamburrini, 2021; Blauth, 2023; Hughes, 2020). This foundational understanding sets the stage for an exploration of the operational and strategic motivations for pursuing such systems, a theme examined in the following section. Building on this conceptual foundation, the next section analyses the practical and strategic motivations that drive the RNLA's adoption of AWS, illuminating the distinct operational challenges and ethical concerns that distinguish autonomy from traditional unmanned systems.

Unmanned systems, though effective in relatively permissive or uncontested environments, are inherently dependent on the ability of human operators to maintain constant control via secure communication links. As a result, these systems are highly vulnerable to electronic warfare, deliberate jamming, and operational disruptions, a vulnerability widely recognized by the interviewees. As one technical officer succinctly stated: *"Remote controlled? Well, that is also difficult, because you have to transmit 4K images through the air. That makes you vulnerable, and connections are often poor in operational areas*." The growing sophistication of adversaries in disrupting or degrading such links not only undermines the operational value of unmanned systems but also amplifies the need for alternatives capable of functioning in degraded or denied environments. Autonomous weapon systems, by contrast, are prized for their ability to operate independently after activation, reducing reliance on potentially compromised communication and enabling forces to maintain operational tempo even in complex, high-threat scenarios. The empirical data from interviews make it clear that two primary motivations underpin the RNLA's pursuit of autonomy:

"Increase combat power. Increase personnel safety. Because those are the two most desired factors." In high-intensity operations, where speed, scale, and dispersion can quickly outstrip the capacity of human decision-makers, autonomy is viewed as a necessary adaptation rather than a discretionary enhancement. As one military programmer explained, autonomy becomes essential: *"in electronic warfare, in situations with high risk for soldiers, and in situations where many tasks have to be performed over a large area with a limited number of available personnel... these are the factors that make autonomy necessary."* The pressures to adopt AWS are compounded by strategic developments on the international stage. The proliferation of both unmanned and autonomous systems among state actors, such as Russia and China, as well as irregular forces, has created a competitive dynamic often described as an arms race. As one officer remarked: *"If we face Russia and do nothing, we will be at a significant disadvantage."* The relative ease with which commercial technologies can be adapted for military use further intensifies this trend: "*You can modify drones cheaply, build them yourself, the technology is widely available."* Operational necessity is accompanied by an ethical rationale, particularly the imperative to protect personnel and avoid technological inferiority that could lead to disproportionate harm. For some interviewees, failing to pursue autonomy is not just strategically risky but ethically questionable. One expert described this sentiment rather straightforward*: "If you don't do it yourself, then you also have an unethical situation. Because then you'll be completely slaughtered."* The decision to pursue autonomy is thus highly context-dependent and continuously reassessed based on operational needs. As another interviewee put it: *"In a context that is not suited for autonomy at all, it is not necessary to make the system autonomous... So why would one choose autonomy? […] Generally speaking, an autonomous system, no matter the degree of autonomy, will be more efficient when information can be shared with the human. Because it is a cooperation between the machine and the commander, who delegates some tasks to the system"* When electronic warfare, time pressure, or operational complexity exceed the limits of human and remotely piloted solutions, autonomy becomes a necessity for mission success (Bondar, 2025; Kwik, 2024)

In sum, the RNLA's pursuit of autonomous weapon systems is a response to both the vulnerabilities inherent in unmanned systems and the shifting demands of contemporary warfare. Strategic competition, operational resilience, personnel constraints, and ethical considerations converge to drive the move from traditional remote-controlled platforms to truly autonomous capabilities. While the motivations for adopting AWS within the RNLA are strong, empirical findings make clear that the integration of autonomy into military operations remains largely at the stage of exploration and experimentation. Autonomy is not yet routinely employed on the battlefield; rather, the RNLA is actively investigating and piloting the conditions and functions under which autonomous capabilities could provide operational value. Within this context of ongoing experimentation, the interview data converge on four principal domains where autonomy could play a transformative role: intelligence and monitoring, autonomous

navigation, including “last mile” operations, autonomous targeting, and swarming. Each of these domains is the subject of current research, prototyping, and concept development within the RNLA, rather than established operational reality.

A foundational area of exploration is intelligence, surveillance, and reconnaissance (ISR). The use of UAS to extend situational awareness is already well established, but the Army is now examining how greater autonomy in these systems could further enhance real-time decision-making. A high-ranking officer, tasked with implementing UAS within the RNLA describes: “*The UAS, in what we currently call an observe role. You use the UAS to look deeper, to see what is happening there, what the opponent is doing, and ultimately that gives you more reaction time.”* While this role remains predominantly under human control, ongoing research and experimentation seeks to determine under what circumstances, and to what degree, ISR functions could be safely and effectively delegated to autonomous systems. The literature notes that ISR remains the “gateway” for early military AI adoption, given the relatively lower ethical risk and higher operational payoff (Verbruggen & Boulanin, 2017a).

Autonomous navigation has become a central focus of research and experimentation within the RNLA as it explores the future potential of military autonomy. Interviewees consistently highlight that autonomous flying presents far fewer technical hurdles than autonomous driving, primarily due to the relative simplicity of the air domain compared to the complex and obstacle-rich environments encountered on land. As one officer remarked: *“UAVs will be used sooner, because autonomous flying is much easier than autonomous driving”.* In the air, the path is relatively unobstructed, whereas ground vehicles are confronted by waterways, trees, and unpredictable terrain: *“You already see autonomous systems flying in the air... Life on the ground is much more complicated. You’re bound to the ground. You have waterways, trees, bushes.”* For military ground vehicles, these difficulties are further intensified by operational constraints such as unreliable GPS signals and the ever-present threat of enemy interference. One respondent explained, *“Autonomous navigation is technically extremely difficult... On the highway, the AI has GPS. Now imagine telling it to go into the forest, without GPS. Autonomous navigation for the military is extremely difficult.”* Despite these significant technical barriers, the RNLA is actively monitoring “last mile” autonomous operations. “Last Mile Autonomous Delivery” describes a scenario in which a drone or loitering munition autonomously completes the final approach to its target, thereby overcoming the effects of enemy electronic warfare and countermeasures. This capability is increasingly relevant in contemporary conflicts such as Ukraine, where such tactics are becoming common practice: *“What you see in Ukraine is what they call, for example, Last Mile Autonomous Delivery. This mainly concerns strike drones. These drones fly toward a specific target from a distance and can autonomously fly the last stretch to the target and detonate there. This is mainly to bypass counter-UAS systems,*

especially electronic warfare means." This function is especially functional for the RNLA, as last-mile delivery actively suits the operational needs with electronic warfare environments. At the same time, experts warn that the progressive extension of autonomy beyond the last mile may blur ethical boundaries and challenge established legal frameworks. As legal policy maker reflected: *"Now it's the last mile. The next manufacturer will say... it's the last two miles. The one after that... the last three miles. So, at what point do we say, this is no longer acceptable?".* This evolving boundary between human and machine control illustrates not only the technical and legal challenges but also the ongoing, and often contentious, deliberation over how much autonomy is needed, and how much is acceptable in modern military operations.

Swarming represents a particularly dynamic and experimental frontier in the RNLA's exploration of autonomy. Unlike traditional single-platform operations, swarming involves the coordinated use of multiple unmanned systems, often drones, to achieve effects that exceed what individual systems could accomplish alone. As one high-ranking officer explained, *"For instance, an effect such as swarming is something we are following. You would have multiple drones, a large group of drones, all in the air at the same time, who can generate a specific effect. That contains a degree of autonomy too."* Swarming offers the possibility of overwhelming enemy defences, enabling distributed reconnaissance, and increasing operational resilience through redundancy. Interviewees noted, however, that the technical and organizational challenges associated with effective swarming remain significant. As one respondent observed*: "You see those drone shows, you see huge swarms of drones all at once… The steps in this are hard to predict. At the moment, it's a factual situation that we don't yet have this kind of technology stable enough to deploy in this way. If it were possible, we would have seen it in Ukraine by now."* While the RNLA is actively experimenting with concepts in which a single operator can manage multiple drones, full swarming remains a future objective rather than an operational reality. Nonetheless, swarming exemplifies how autonomy can radically reshape tactical possibilities and simultaneously generate new and complex ethical dilemmas concerning distributed control and responsibility. From an ethical perspective, swarming raises important questions about the dilution of human agency and accountability in decision-making. As control shifts from direct human oversight of individual platforms to supervisory management of autonomous collectives, the risk emerges that critical decisions, such as target selection, engagement, and escalation, may become increasingly opaque and difficult to attribute to specific individuals. This "problem of many hands" complicates the assignment of responsibility and undermines established norms of military accountability(Santoni de Sio & Van den Hoven, 2018). The rapid, adaptive behaviours enabled by swarming may further strain the ability of human operators to meaningfully intervene or override system actions in real time, challenging the very notion of "meaningful human control."

In contrast to domains such as navigation or swarming, the application of autonomy to targeting and firing decisions is subject to far greater ethical and legal constraint. Interviewees consistently emphasized that, while technological capabilities for autonomous targeting are advancing rapidly, the delegation of lethal force remains fundamentally distinct from other operational tasks. As one technical specialist explained, *"Autonomous shooting. Technically, that is not difficult... but I can imagine that you can raise some ethical objections here. The use of lethal force is an enormous responsibility. You cannot place that… at least not yet… with autonomous systems."* This heightened caution is grounded in the centrality of the targeting cycle within Dutch military practice, a procedural safeguard institutionalizing the requirements of proportionality, distinction, and necessity under IHL. Unlike navigation or intelligence tasks, targeting requires context-sensitive judgment and nuanced moral reasoning that directly affect civilian protection and the legitimacy of military action. Accordingly, both policy and practice widely regard meaningful human control as indispensable for decisions involving life and death. Practitioners largely agree that, while autonomy can enhance many military functions, the ethical threshold for targeting is far higher, demanding rigorous oversight, reflection, and accountability. These judgments are not merely procedural but fundamentally moral, relying on context, subjective experience, and empathy, qualities that current and foreseeable AWS lack. Nevertheless, some proponents of AWS contend that not every aspect of the targeting cycle must be delegated to machines; rather, select tasks could be entrusted to autonomous systems under strict conditions. The central challenge, then, lies in determining which elements, if any, can ethically and legally be delegated, and under what constraints. As a former military officer now active in the drone industry explained, clear frameworks and operational boundaries are crucial*: "Usually you proceed: Yes, it's necessary, yes, it's applicable. Then we make a specific plan. Which actions that specific system must take, including under which framework. For example, authorizations to do something in an area. For example, use of weapons, outside that, not. Specific time, within this time, yes; outside that, no. For these specific targets, yes; others, no. With a requirement for a prior report, for example, and a human approval. This is how the commander, supported by his specialists, builds the framework. A sort of ring of obligations, within which authorization for autonomous decisions can take place."* Yet, even with robust procedures, delegating certain targeting tasks to AWS can jeopardize the integrity of the entire targeting cycle if the system errors. Thus, the very aim of the targeting cycle, to safeguard IHL compliance, comes to depend on the reliability of the autonomous system in performing designated functions. Recognizing this, some operational experts advocate institutional safeguards, such as protocols that require AWS to default to inaction and escalate uncertain situations to human operators. As one former commander now working in industry described: *"Yes, it is also true that at a certain point the autonomous system ends up in a situation where it does not have sufficient confidence for an identification. That leads to a non-attack situation. And it could refer*

that situation to a human. Yes. Assuming there is a connection. Yes. Then it is human judgment whether it could be an attack, an override of the autonomy, or by a non-attack. That is possible. So, in case of doubt, consult the human, basically. Yes." However, the effectiveness of such a safeguard depends fundamentally on the AWS's capacity to recognize the limits of its own certainty, a capability that is far from guaranteed, especially in complex or ambiguous environments. As another respondent warned: *"What if the system makes a mistake with identification? Some situations are already difficult for humans to make proper identification, how should a robot do that? You really have to keep Human-in-the-Loop, because the system simply cannot identify well enough and sometimes humans can't either."*

This is why man-machine cooperation must be managed with the utmost rigor: the commander not only carries ultimate authority but also bears the unique burden of translating legal and ethical principles into operational practice. It is essential that the commander has absolute trust in the autonomous system's reliability, transparency, and capacity for clear information-sharing. Only with this trust can the commander make informed judgments about when, where, and how to delegate autonomy, knowing that his instructions will be understood and his boundaries respected. As one military expert explains: *"Conceptually, autonomy is directly linked to the person who holds the mandate. That person decides whether or not to delegate autonomy to the machine. ... The commander of the military unit, who is chosen and trained, and familiar with his system, has the obligation to assess necessity and applicability."* This obligation is not merely theoretical but embedded in daily military routines. The commander's ongoing risk analysis must encompass not just operational necessity and feasibility, but also the technical maturity and contextual appropriateness of the autonomous system: *"He must assess whether there is a necessity for forms of autonomy. And which forms of autonomy? In this sense, it is a flexible system. He has the necessity and obligation to assess the applicability of autonomy in the context he is considering."* Even highly autonomous systems are expected to maintain channels for information-sharing, so the human commander can sustain situational awareness, intervene when necessary, and fulfil accountability requirements*: "Because it is a cooperation between a commander who wants something, who gives instructions, and a system that must execute. Even if it is highly autonomous, it is better if it can share information, so the commander can assess, form his picture, and make decisions."* Yet, as previous sections have shown, this trust cannot be assumed. Direct and reliable communication between human and machine is not always guaranteed, especially in high-tempo or denied environments. For this reason, the practical implementation of man-machine cooperation requires robust pre-mission risk mitigation, meticulous configuration of the autonomous system, and a clearly defined set of permissible actions and safety boundaries. As another interviewee outlines: *"You first build the framework. And only then the permission. ... And that is of course very context dependent. In my hub I configure the agents. Including a number of safety checks, especially regarding ammunition and weapon use. In my risk*

analysis I go through a number of potential risks. For example, is this algorithm trained and validated to distinguish these types of targets in this context?"

Building on this foundation, Dutch policy and operational practice now increasingly stress the principle of context-appropriate human control. The degree and nature of oversight must be dynamically tailored to both the level of autonomy and the complexity of the operational environment. As one legal policymaker explains: *"The simplest way is a little diagram that NAME once made. To visualize it very simply: an x-axis and a y-axis. One axis increases in degree of autonomy... The other in operational complexity. So, the more autonomous the system in a more complex operational environment, the 'more dangerous' it becomes, and thus, the stricter it should be regulated."* This framework makes clear that autonomy is not a binary attribute but a spectrum, requiring regulatory and operational safeguards to scale accordingly. At one end, systems like traditional firearms are fully human-operated; at the other, highly automated defensive systems like the Goalkeeper or Phalanx may be allowed substantial autonomy in low-risk environments such as open sea. *"Because it is at sea. And at sea the number of civilians is limited, as is the risk to victims... you can say, the chance of something going wrong toward civilians or civilian objects is so small that such a system could reasonably operate autonomously."* However, once such systems are deployed in more complex settings, such as land-based air defence in densely populated areas, the calculus shifts. As the interviewee notes: "*That's the environment, so the operational context is much more complicated... There are plenty of examples where a Patriot system, for one reason or another, suddenly targeted a commercial Boeing... So, these are all things to take into account. Then the operational context is different. There are more risks for civilians and civilian casualties."* In such environments, the system's ability to distinguish civilian from military objects is inherently limited, demanding tighter human supervision. Conversely, if the context is tightly controlled and the risk to civilians is negligible, as may be the case in certain conflict zones, greater autonomy could be justified: *"But if it's a situation where, for example, Ukraine-Russia, as is happening a lot now, there are simply no civilians in this area... The chance that a civilian is present is 0.001%. Then I don't really care that the system can't see that. It can just do its job."*

Ultimately, the human control process centres on the commander's deliberate decision whether or not to deploy the autonomous system in a given context. This decision is far from automatic as it demands a careful weighing of technological reliability, operational necessity, and the risk environment. The level of human involvement, whether "prior to the loop" (through careful pre-mission configuration and rule-setting) or "in the loop" (with real-time intervention and veto power), is thus set by the commander, who is positioned to judge how much oversight is required: *"And then the question is, how much human control do we find acceptable to deploy such a system? That depends on what you do prior to the loop, so*

what parameters you set. How well can we configure it so that it actually finds the right target and doesn't make mistakes, for example. And how well can it understand its environment." By making these choices explicit and context-sensitive, the commander not only ensures that operational and ethical standards are met but also secures the accountability process, anchoring responsibility in a clearly identifiable human actor. This preserves the principle that, even in an era of increasing autonomy, there remains a person who can be held responsible for both the decision to delegate and the outcomes that result. This central role of the commander in (pre-)determining the extent and nature of human involvement in autonomous operations underscores a critical point as accountability and responsible command hinge not just on the existence of oversight, but on the commander's ability to foresee and manage the system's behaviour. The problem of foreseeability is fundamental for legal and moral responsibility (Bovens, 2007). As one expert observes: *"If you're going to, for instance, hold somebody accountable, there has to be a level of predictability in that it was going to go wrong. You can't hold somebody accountable for something that they could have in no way known was going to happen.*" In other words, for accountability to remain meaningful, commanders must have both the information, and the confidence needed to anticipate how an AWS will perform under specific conditions. If the autonomous system's actions are unpredictable or its failures unforeseeable, holding the commander responsible becomes untenable. Thus, the commander's decision to deploy an AWS is inseparable from the requirement that the system's behaviour be both reliable and predictable, qualities that are prerequisites for justifying the delegation of control and for upholding legal standards. Predictability is not just a technical ideal but a necessary condition for responsible military conduct and enforceable accountability. Ultimately, the ability to anticipate and trust an AWS's behaviour depends on how the system is designed, validated, and tested in practice. In the following theme, I will examine how requirements for predictability and reliability are addressed at the technical level.

4.3 Theme 3: Technical Investigation

The technical architecture underpinning AWS is central to their predictability, reliability, and accountability. As established in the previous theme, a commander's ability to responsibly delegate decision-making to autonomous weapon systems depends not only on procedural safeguards and ethical considerations, but fundamentally on the inner workings of the technology itself. The degree to which an AWS can be trusted to act within defined parameters, and the extent to which its behaviour can be anticipated, explained, and scrutinized, are direct consequences of how it is designed, programmed, and tested. In other words, the technical architecture does not simply enable autonomy; it determines whether autonomy can be exercised in a way that is transparent, controllable, and compatible with the requirements of international humanitarian law. This theme examines the core technical features and

safeguards that are necessary to support meaningful human control and ensure that AWS can operate predictably and reliably in diverse and complex operational environments.

Technical decision making

At the heart of every autonomous weapon system lies a decision-making framework that governs how the system perceives, processes, and responds to its environment. The architecture of these frameworks shapes not only the system's operational effectiveness but also its transparency, predictability, and accountability. In military applications, the design of decision-making systems must reconcile the need for precise, controllable actions with the demands of complex, uncertain, and adversarial environments. Understanding the foundations of these systems is crucial for evaluating both their capabilities and their limitations. The interviews reveal a distinct division between older, rule-based approaches, such as preconditional programming, and newer, data-driven models based on machine learning. Understanding the foundations of these systems is crucial for evaluating both their capabilities and their limitations. The most fundamental, and still the most prevalent, approach to programming autonomous behaviour in military contexts is known as preconditional or "IF-THEN" programming. This method relies on explicit, pre-defined rules that instruct the system how to act in specific circumstances: IF a particular condition is met, THEN the system executes a predetermined response. For example, an autonomous vehicle may be programmed to stop IF it detects an obstacle in its path, or to alert an operator IF a potential target enters a monitored zone. Behaviour trees are a related programming technique, structuring multiple IF-THEN rules into a logical hierarchy to handle more complex sequences of action and decision-making. This rule-based logic maps directly onto military thinking and doctrine, where procedures and tactical responses are carefully codified. As one military programmer explains: *"With an if-statement, you know: at a certain threshold, if you set it, then a decision is made. And you can check that afterwards using images."* The alignment between software engineering and tactical doctrine is clear: *"Military thinking works very procedurally. We have TTPs, tactics, techniques and procedures, which all soldiers learn. If this happens? Then you do that. Contact front? Reaction X. You learn that from day one. And the same applies to vehicles, especially at the lowest level. The beauty is that you can also program this behaviour, which is what it is, and that's then very much based on if-statements."* Preconditional programming offers two principal advantages: transparency and traceability (Christie et al., 2024; Santoni de Sio & Van den Hoven, 2018). Unlike human operators, whose decisions may rest on intuition or incomplete recollection, a system programmed with IF-THEN logic leaves a clear, auditable record: *"You can therefore review afterwards why a decision was made. And I actually think that is even more ethical than how a human does it now."* In practice, after-action reviews can be conducted using systematically recorded data: *"A human does everything based on experience: 'I thought it was an enemy vehicle, so I shot.' And*

afterwards, it turns out it wasn't. But how do you check that? Often, there aren't even camera images." This is an advantage in the enhancement of accountability. Because every system action is justified by explicit thresholds and conditions, responsibility for the programming, and for operational mistakes, can be more precisely traced. Codifying decision logic provides both operators and commanders with a robust framework for post hoc scrutiny and learning. Here it is important to notice that this is not strict logging of outcome but also logging of how these outcomes have come to place. However, this predictability also introduces vulnerabilities. If adversaries can infer or reconstruct the underlying conditional logic, they may deliberately exploit it. "*If the enemy knows exactly how your procedures work, they can manipulate them. They know exactly what action to take to trigger your system."* Thus, while the system's behaviour is fully defined and controllable, this very determinism can make it susceptible to adversarial countermeasures. Moreover, there are significant practical challenges in designing preconditional systems. Every possible scenario and dilemma must be foreseen and codified in advance, a difficult, if not impossible, task in the complex and dynamic context of modern warfare. As one interviewee notes: "*The big problem with what-if systems is that you have to specify everything in advance. Every dilemma must be explicit."* Rare, ambiguous, or novel situations might be missed, and the system could behave unpredictably or inappropriately if it encounters something for which it was not programmed. Here it is possible code that the system aborts missions with unclear rules, this however lowers its operational efficiency. The difficulty is not just in specifying actions, but also in establishing the right thresholds for those actions: *"The hardest part is determining the correct thresholds. For example: when may the system fire? Perhaps at a recognition score of 97%? Or only at 99%?"* The decision of where to set these thresholds has direct operational and ethical consequences, as too low a threshold risks false positives, while too high a threshold may cause hesitation or missed opportunities. Threshold-setting thus becomes a balancing act between operational necessity, technical reliability, and ethical responsibility.

In sum, IF-THEN and rule-based programming remain at the core of contemporary military autonomy due to their transparency, compatibility with military doctrine, and auditability. Yet, their limitations, especially in dealing with ambiguity, unpredictability, and adversarial deception, underscore the challenges of translating the complexity of real-world warfare into static procedural code. This rigidity in unforeseen circumstances has motivated experimentation with more flexible, data-driven approaches, most notably, machine learning and neural networks. These methods differ fundamentally from rule-based programming: rather than following pre-programmed rules, machine learning systems are trained on large datasets to identify statistical patterns and correlations (ICRC, 2019; Kwik, 2024). In principle, this enables systems to generalize from previous experiences and adapt to new inputs, even when those have not been specifically anticipated by programmers. Machine learning and neural networks have therefore become especially attractive in the military context due to their promise of adaptive and flexible

decision-making, which is difficult to achieve with traditional rule-based programming. Unlike preconditional systems, which require all possible scenarios to be manually encoded, machine learning models are designed to learn from large amounts of operational data and potentially adapt to new, unforeseen situations. This potential for flexibility and real-time adaptation is highly valued in the complex and dynamic environments where autonomous military systems are being developed. Machine learning models are also more adept in synthesizing and analysing large amounts of data. As a policy maker explained: "*There is a kind of gradation in this. If you move away from classic systems, it's partly about how you get a coherent world view. That's quite easy for a system like Goalkeeper with a radar but if you use, for example, video material, or Lidar, or whatever, and you have to integrate all that input, then you have a whole mountain of data, and you have to make sense of it. That's when machine learning comes into play; that's already an extra layer of complexity. Plus, you also bring more of that into the decision-making process."* This illustrates how machine learning, in theory, enables systems to fuse and interpret vast, heterogeneous streams of sensor data, creating a more "coherent world view" in situations where simple rules are insufficient. In practice, this means that machine learning could help autonomous platforms manage and respond to highly variable operational environments, potentially increasing speed, reducing the cognitive load on human operators, and enabling mission continuity even when communication with humans is limited or lost. The ambition, as described in both expert interviews and the literature (ICRC, 2019; Kwik, 2024; Schwarz, 2021), is to build military systems that are not merely reactive, but genuinely adaptive, able to learn from experience and continuously improve their performance in real time. However, this flexibility comes at a cost. Unlike IF-THEN programming, the inner workings of machine learning models, especially deep neural networks, are typically opaque, even to their developers. This "black box" nature means that, while the system may deliver correct results in many scenarios, its decision-making process is difficult to interpret or audit after the fact. As one military programmer cautions: *"With a neural network, you might get a slightly better outcome at that moment, but there's a huge danger: the system might do something you absolutely don't want, and then you lose control."* This lack of transparency has profound implications for accountability, traceability, and ethical oversight, challenges widely discussed in both policy and academic debates (Christie et al., 2024; Santoni de Sio & Van den Hoven, 2018)

Another, yet connected, technical limitation of machine learning is its reliance on non-causal reasoning. Rather than understanding the "why" behind observed relationships, ML systems excel at identifying correlations in historical data, without genuine comprehension of cause and effect. As the programmer explains, *"All these beautiful AI systems, they are non-causal systems. They are good at identifying patterns, but not at cause-and-effect reasoning."* This limits their utility in tactical or ethical decision-making, where nuanced judgment and intuition are often required. *"Especially in these kinds of military,*

creative scenarios... then you need a human with intuition. Who can estimate: if someone is walking with aluminium foil, there's probably an enemy soldier behind it. But even that is dangerous. Maybe it's actually a civilian behind the robot. Only a well-trained human, who has already experienced such a scenario, can judge this correctly." Another concern is the unpredictable behaviour of machine learning systems in novel or adversarial contexts, precisely the situations in which military operations frequently occur. As one expert on AI in land warfare remarks: "*The whole point of combat is to do the unexpected. The whole point is surprise, to gain initiative, do something surprising, do something you haven't done before, do it in a way that hasn't been done before. And the whole thing about AI systems is they're programmed off historical data. So, the chances of you knowing that it was going to behave like that in a novel situation is very slim."* Even high-performing neural networks can make critical errors if they encounter situations outside their training data, a risk compounded by adversaries who may deliberately introduce unfamiliar scenarios. This opacity and unpredictability also challenge post-hoc review and accountability. *"If you train an AI to recognize certain items, and you do that with a neural network, you can't explain it. You don't know exactly how it arrived at the conclusion. And I wouldn't want to be the technician who trains an AI on material recognition, only to find out that a school bus has the same headlights as a certain tank model. And then it says, well, that's a tank, I'll fire a missile at it. That's why I think we should be cautious with such things."* Despite rapid advancements in artificial intelligence, a profound gap persists between the reasoning capabilities of machines and those of humans, especially in situations requiring creative, ethical, or ambiguous judgment. As one programmer notes: "*I just don't believe in it yet. AGI still can't do causal reasoning. That's why it will never be better than maybe the 'dumbest' human, or rather, the average human. Because they can already reason quite well."* Until such breakthroughs in causal reasoning occur, neural networks and machine learning systems, while powerful tools, remain fundamentally limited by their non-causal, opaque nature. This restricts their suitability for fully autonomous use in lethal or highly ethical military decisions, as both the literature and expert interviews make clear.

Yet, these technical limitations do not mean that machine learning is without value in military operations. On the contrary, while machine learning may not be practical for lower-level autonomous decision-making, it has proven especially valuable for processing and filtering the vast streams of sensor input now available on the modern battlefield. As one interviewee reflects: "*Not for the individual UGV, no. Perhaps later for larger formations, such as a battalion or platoon, when coordinating missions. There, the individual decisions matter less, because the vehicles themselves still have a decision tree. The behaviour of hundreds or thousands of vehicles is then determined by the network, but that is more about efficiency and coordination."* When asked where neural networks are actually used, the answer is clear: *"For image recognition or terrain analysis, the input side. So, you have a camera, the system sees*

something, and that input must be processed. But the translation from 'I see something' to 'should I fire or not', that decision-making must be procedural." Thus, while algorithms are essential for automating and structuring the deluge of data modern sensors generate, they remain tools for supporting, not replacing, critical, lower-level, and especially lethal, decisions. Most machine learning models used for these tasks are based on neural networks trained on massive datasets consisting of visual input, pixels from cameras, thermal signatures, or Lidar data. As one interviewee explains: *"Object detection model. These are the well-known trained neural networks, used for visual identification based on, among other things, pixels and, for example, a series of consecutive observations. There are a number of best practices, and it can work well if done properly."* Reliability in object detection does not depend solely on the volume of data, but also on the effective integration of various sources and types of input. Effective deployment requires algorithms not only to process raw data but also to apply procedural rules, such as temporal and spatial filtering, to minimize mistakes and improve reliability. *"It works better if, for example, you require at least five different observations from different angles. That already eliminates a lot of false positives. So that's the temporal factor. Another best practice is to use the camera's location data to determine the size of an object."* By combining these sensor fusion strategies, AWS can achieve a higher degree of confidence in their outputs, but they are never completely infallible.

Confidence

A hallmark of machine learning systems, and a challenge for their adoption in high-stakes military environments, is their ability to generate quantitative outputs in the form of confidence scores. This score is a numerical value, typically between 0 and 1, that expresses the model's estimated certainty that a detected object is, for instance, a tank or a truck. As an interviewee observes: *"The bizarre thing is, I think an autonomous system will come up with a number sooner than a human. With machine learning, especially neural networks, you usually get a bounding box with a 'confidence score'. But what does that number mean? That still needs to be researched. People just say, 'that is an enemy' or 'I don't know'. An autonomous system comes up with a number, but how do you interpret that?"* The use of such scores brings both opportunities and new challenges. In theory, it enables military operators to set and enforce operational thresholds, for example, "only engage if confidence > 0.97". However, the interpretation and trustworthiness of these scores are not always clear, especially as even well-trained models can fail in unfamiliar contexts. Furthermore, the relationship between a confidence score and the actual risk of misidentification remains difficult to calibrate and explain, highlighting the limits of current explainability and the transparency problem in modern AI (Hughes, 2020; Kwik, 2024; Longpre et al., 2022). There are no universally accepted, quantifiable metrics for comparing machine and human performance in tasks like target identification. Human decision-making in the field is fundamentally

qualitative, rooted in intuition and accumulated experience: *"What I really miss is how we quantify people's ability to identify enemy vehicles. There are no metrics for that. It's all based on experience, looking at a lot of pictures, being in the field, counting road wheels, looking at the barrel. That's how you reach a judgment."* Setting "acceptable" accuracy norms, such as a 95% or 99% confidence threshold, thus becomes as much a matter of policy and ethics as of technical specification: *"The hardest part is determining the correct thresholds. For example: when may the system fire? Perhaps at a recognition score of 97%? Or only at 99%?"* To bridge this gap, technical tools such as the confusion matrix are increasingly central to assessing the performance of object detection models. The confusion matrix tracks the number of true positives, false positives, false negatives, and true negatives, providing a structured framework for understanding both successes and failures: *"Not only does it identify the right target. But also, what the situation is for false positives and false negatives. You really want to get your false positives as close to zero as possible. Something that is not a target but is identified by the neural network as a target. You really want to get those as close to zero as possible. Certainly. You use your data and tooling for that."* Despite these tools and best practices, it is widely recognized by both practitioners and scholars that perfect accuracy remains unattainable: *"And what kind of accuracy are you aiming for? That is hard to express. Achieving 100% is technically and practically not feasible. And suppose we express it as a confidence level. 0.95 out of 0 to 1. That sounds pretty high. But what does that mean? What does it actually entail? How is it established? It certainly means something. 0.95 is higher than 0.6. And 0.3 doesn't sound good. We'd better not do that. It certainly means something, yes. But what does it actually say? I think that is more important. And I also learned from a technology project to work with that confusion matrix. That I can say with 0.99 certainty that my score on false positives is good, virtually zero. So, it is impossible to express yourself in absolute certainty. That is part of risk acceptance.* Ultimately, the organizational responsibility is to ensure responsible use through best practices, thorough testing, and transparent risk acceptance: "*That is also the obligation the organization has: responsible use of this technology. I cannot guarantee that it works one hundred percent or not at all. But I can organize everything: best practices and thorough testing, evaluation, and verification. That I deal with it organizationally responsibly. So that the soldier in the field can rely on it. That can be done. Absolutely."*

In sum, the problem of quantifying and interpreting recognition in autonomous systems is deeply intertwined with technical, ethical, and operational questions. It requires not only advances in algorithmic design, but also ongoing organizational attention to best practices, testing, and the setting of defensible policy thresholds, always with the recognition that absolute certainty is unattainable and risk can never be fully eliminated.

Data Management

The operational performance and reliability of machine learning-based autonomous systems are inextricably linked to the quality, quantity, and relevance of the data they are trained on. As articulated in NATO's revised AI strategy: *"The availability and management of AI-ready, quality data is a prerequisite for the development and use of secure, reliable and responsible AI systems. Quality data is foundational to the development of effective AI-enabled systems since all analytic and AI capabilities require trusted, quality data, which does not include unintended bias, to support the development of such systems"* (NATO, 2024). This is echoed by the interviewees: "*If you set high standards for operational performance... you automatically set high standards for your data. Data quality and your data process, including implementing TEVV [Testing, Evaluation, Validation, Verification]. Otherwise, you are not acting responsibly. So, you set high standards for your data, which, for example, must be representative of the context and the performance that will be required in operation."*

A recurring theme among experts is the critical need for congruency between the data used to train machine learning models and the real-world operational environments in which autonomous systems will be deployed. This congruency, often referred to as "domain alignment," is essential because machine learning models do not possess genuine understanding or intuition (Amoroso & Tamburrini, 2021; Christie et al., 2024; Kwik, 2024). Instead, their ability to perform tasks, such as recognizing targets or distinguishing between civilian and military objects, depends almost entirely on the patterns and features present in the training data. If the operational environment presents situations, visual cues, or types of objects that differ from those encountered during training, the system may fail catastrophically. As one interviewee illustrates, "*If you train an AI on a vehicle that is always in the forest, and then you see it in the city, you might conclude: this is not that type of vehicle, because it belongs in the forest. I am in the city, so it cannot be it."* This example underscores a fundamental limitation: machine learning systems extrapolate poorly outside of their learned context, making them unreliable when faced with novelty. This challenge is further exacerbated by the complexity and unpredictability of military operations, especially on land, where the environment is highly variable. Unlike air defence, where *"the airspace is fairly well defined,"* the land domain is filled with factors that can affect the performance of recognition and decision systems: varying terrain, the presence of civilians, changing weather, camouflage, and a wide range of vehicles and equipment. *"On land, there are many more factors. Think of civilians, own troops... So, the land domain is simply more complex. And that increases the margin of error."* From an ethical and legal standpoint, this lack of congruency between training and deployment environments also poses major risks for compliance with IHL. Systems trained in narrow, idealized contexts may fail to make critical distinctions in the real world, potentially misidentifying civilians as combatants, or failing to

recognize protected objects. As a result, the reliability and lawfulness of autonomous weapon systems hinge not only on their underlying algorithms but, critically, on the realism and breadth of their training data. Achieving this congruency requires both technical investment in simulation, data collection, and ongoing retraining and careful organizational attention to the realities of warfare. To mitigate these limitations, practitioners increasingly rely on synthetic data generation and simulation to broaden the range of operational conditions the model is exposed to. As one interviewee explains: *"For material recognition, we use a program with synthetic data generation... These are placed in the world, running in different environments, with fog, with rain, day, night, thermal imaging, with and without camouflage, fully or partially visible. So, we try to eliminate as many possible deviations from the system as possible."* Still, even sophisticated simulations cannot fully capture the variability of real warfare, including human behaviour, emotional responses, and adversarial tactics: *"The adrenaline, all the feelings and changing behaviours that come with it, are very difficult to simulate."* The theoretical possibility exists to train for a wide range of contexts, but practical and financial constraints are immense. As one exchange highlights: *"Is it even possible to make all these kinds of contexts and these potential scenarios into the data, into training data?" - "I think it's possible, theoretically... but at what cost and what are the trade-offs on other things? I think it's expensive, I think it's time-consuming."* Compounding this challenge is the dynamic nature of military conflict, adversaries adapt their behaviour in response to deployed systems: "*You have to have it adaptable, because once your opponent learns that you've built a system that is only based off of the ability to detect a tank, then they're going to stop dropping tanks."* This underscores the necessity of continual model retraining and adaptation, as well as robust testing across a wide range of scenarios.

Best practices in machine learning development emphasize the importance of maintaining a rigorous separation between training and testing data. This principle is foundational in the field of artificial intelligence, as it helps prevent overfitting, a situation where an algorithm learns the specifics of its training examples too closely, resulting in high performance on seen data but poor generalizability to new, unseen scenarios (Kwik, 2024). Especially for high-risk and safety-critical military applications, ensuring that models are not just memorizing but are actually able to generalize is crucial. As one respondent states, *"For high-risk applications... your test set is separate from your training set. That is a basic principle."* The purpose of this separation is to simulate the unpredictability and variability of real-world operations. By testing a model on data, it has never seen during training, developers can better estimate how the system will perform under actual battlefield conditions, where surprises and edge cases are common. This approach is meant to surface failures or blind spots before deployment, providing a measure of confidence that the system's behaviour will be reliable and predictable in the field, not just in controlled lab environments. However, as emphasized by interviewees, the RNLA currently lacks

sufficient data to implement this principle to the required standard. As one respondent notes, "*We have so little data that we don't have enough to do a full training, and then a variety of testing. So that is something that would have to be part of a data strategy."* This shortage is not simply an inconvenience, it is a structural limitation that fundamentally constrains the ability to train, test, and validate machine learning models for autonomous weapon systems. In particular, rare, novel, or high-risk scenarios, such as the identification of new enemy vehicles, complex terrain, or evolving adversarial tactics, are extremely difficult to represent in available datasets. Beyond the technical necessity, there is a critical legal dimension, which can complicate the article 36 procedure. Under IHL, any weapon system must be proven to comply with the principle of distinction, being able to reliably differentiate between legitimate military targets and civilians or protected objects. As one expert points out, "*Under IHL, for a weapon to be legal, it has to be proven that it can do distinction. How you do that using what data set, how we define the data sets, and which ones we say, if it performs on this data set, then we deem it to be distinct."* This highlights the legal and ethical imperative for transparency, traceability, and rigor in the data processes that underpin autonomous weapon systems.

In summary, the separation of training and testing data is not merely a technical formality but a foundational safeguard for both operational reliability and legal compliance. Overcoming practical data challenges, and developing robust data strategies, are thus essential steps for any organization seeking to responsibly deploy machine learning in autonomous weapon systems. Yet, as established in the previous subtheme, the technical performance and reliability of AI-driven autonomous weapon systems depend not only on their algorithms, but fundamentally on the quality and relevance of the data that trains them. The challenges of ensuring high-quality, context-appropriate data extend well beyond technical calibration, presenting major operational, legal, and ethical dilemmas for the RNLA and any military organization seeking to responsibly deploy AI-enabled systems. A persistent concern among policymakers and practitioners is the risk of deploying autonomy based on incomplete or irrelevant data. As a legal policy maker within the Ministry of Defence notes, *"I think it is so critical that, for now, the more autonomy is based on that data, the less likely it is to be deployable, until we deem the data good enough."* The consequences of poor data quality are not abstract: without a robust foundation, autonomous systems may behave unpredictably or even dangerously, especially in the unpredictable environment of high-intensity combat. This was vividly described by one respondent: *"Just in terms of high intensity combat, the speed at which it's happening, the AI would be a little bit too rigid and... provide a risk to soldiers because if the soldiers were reliant on that piece of technology and it failed in the battlefield, then you're essentially either endangering the soldiers, making the soldiers do all the work, or they're not trained on how to perform if that piece of technology is there, so it's not a reliable enough piece of technology."* Thus, ensuring data congruency is a fundamental precondition for both ethical use and mission effectiveness.

Yet, achieving such congruency is far from straightforward. As previously discussed, the RNLA currently lacks sufficient control over relevant, high-quality data required to train and validate autonomous weapon systems.

Data ownership plays a critical role here, as commercial companies are often at the forefront of innovation in AWS development. However, these companies' proprietary interests and differing incentives frequently limit military access to essential data libraries. As an active Ukrainian commander highlights: *"Importance of such systems is very high, but there's lots of barriers. One of them is target libraries. Considering high competition on the market between manufacturers, not everyone is ready to share the data. And also, such libraries can be kept secret and closed for civilians (as most of the developers of unmanned systems are civilian) to use due to bureaucracy."* In this way, commercial logic often stands in direct tension with military operational needs, making it difficult for armed forces to independently verify, update, or adapt the data underlying their autonomous systems. This dependency prompts a crucial debate within the Dutch defence sector: should the RNLA rely on external vendors for critical capabilities such as target recognition, or should it prioritize internal ownership and stewardship of algorithms and data? Many interviewees expressed deep scepticism towards outsourcing, warning of strategic and ethical risks: "*Well, there is another option: you buy it from industry. But I don't believe in that at all. ... If you outsource the decision-making to a company, you are essentially outsourcing warfare. That might be a strategy, but I think that control over those decision-making algorithms must always remain with us."* Such views are echoed throughout the organization: "*Ownership of the data and the algorithms must really belong to Defensie. Otherwise, you lack decisiveness."* The ability to test, understand, and certify systems, essential for legal and operational accountability, can only be guaranteed if the RNLA maintains ultimate control. As one respondent summarizes: *"Even if we don't make it ourselves, we must at least be able to verify if it's correct. Ultimately, Defensie must have a certification authority to determine if such a system should be purchased."* Thus, a critical requirement for the operational reliability and ethical deployment of AI-driven military systems is the continuous management of both data and algorithms throughout their lifecycle. Unlike traditional weapons, which can remain static for years, autonomous weapon systems operate in dynamic environments that constantly evolve, driven by adversarial adaptation, shifting operational requirements, and rapid technological developments. As one subject matter expert stressed: *"You also want to update it during the lifecycle. Because your training set might become outdated."* If data becomes obsolete or unrepresentative, system performance and reliability quickly degrade, threatening both mission effectiveness and legal compliance. To keep pace, the process must begin with the ongoing collection and curation of operational data. This goes beyond digital files, interviewees emphasize the need to identify data gaps and adapt sensor hardware as environments and missions change: *"You need to start identifying gaps in your data and then*

informing the hardware side, which is the sensors, to say, OK, we actually need to get sensors collecting a wider range of data. We need to build new sensors onto our vehicles so that they collect the data that we need." Thus, lifecycle management is not a purely technical routine but a system-wide organizational concern. The retraining and validation of algorithms with newly acquired more contextually relevant data is therefore of immense importance. This is not a one-off event but a recurring organizational responsibility. One interviewee underlined: "*The quality of AI improves with the quality and quantity of data you input into it,"* directly linking continual data management to the reliability of fielded systems. To ensure responsible and lawful deployment, periodic retraining, rigorous validation, and battlefield verification must become embedded in doctrine and practice. Every update to training data or algorithms must be systematically tested and certified, to maintain confidence in performance and compliance with both legal and ethical standards. However, sustaining responsible AI deployment is not just a matter of establishing robust technical systems and processes; the cultivation of internal expertise and an adaptive organizational culture is equally essential. For the RNLA, the effective stewardship of autonomous weapon systems requires personnel who are not only capable end-users, but also skilled programmers and critical thinkers, individuals who can understand, adapt, and improve the underlying technologies as operational needs evolve. As one interviewee emphasized: "*For that, I need soldiers who can program. That is the most, most important thing. Without those people, you would have to buy it from the industry, but I don't believe in that."* This sentiment reflects more than a practical concern, it underscores a fundamental strategic risk: when technical know-how is outsourced to commercial vendors, the military risks becoming dependent on external actors for the most sensitive aspects of warfare. Maintaining persistent lifecycle management within the military is therefore not only a technical imperative but a matter of operational sovereignty and strategic control. If the ability to update, test, and adapt algorithms or data libraries is lost to third-party providers, so too is the capacity to respond swiftly and independently to new threats or changing circumstances. As another respondent clearly articulated: "*My opinion remains that we must build it ourselves. Everything, from doctrines to tactics, must remain in our own hands. We must be able to determine: this platoon carries out this assignment, with this system, at this location."* In this light, investing in in-house technical expertise becomes inseparable from safeguarding national security interests and preserving the autonomy to define, adapt, and execute military operations according to Dutch values and priorities.

In summary, the continuous management of data and algorithms is not an auxiliary technical task but a persistent, multidimensional challenge spanning technology, people, doctrine, and culture. Only through active stewardship, robust internal expertise, and organizational commitment can the Dutch military safeguard both operational readiness and ethical integrity in an era of accelerating technological change.

4.4 Theme 4: Organizational challenges

The integration of artificial intelligence and autonomy into the RNLA demands a fundamental transformation in how military systems are acquired, developed, and deployed. Traditionally, the Dutch military has relied on a linear, static model, one in which platforms such as tanks or artillery are purchased, fielded, and maintained for decades, with only occasional upgrades. However, the accelerating pace of technological innovation, particularly in the fields of AI and unmanned systems, renders this approach increasingly obsolete. Instead, the RNLA must embrace a paradigm of continuous, adaptive development and operational integration, in which systems are persistently improved and tailored to evolving threats and operational requirements. Crucially, this transformation is not solely technological but reaches into the organizational, legal, and ethical fabric of the military. A recurring theme in expert interviews is that autonomy in the military context is not about replacing humans, but rather about aligning the behaviour of machines with the needs and realities of military users. As one project leader for autonomy development emphasized, *"My experience is mainly that, besides being able to program, I always look from the user's perspective. I try to map out very well what the military user wants from autonomy."* This user-centred approach is essential for fostering both operational effectiveness and ethical acceptability in deploying autonomous systems. The centrality of the user is no coincidence. Modern conflict is marked by high complexity, ambiguous situations, and shifting ethical stakes, all of which require not only technical sophistication but also a clear understanding of what autonomy is for, and what it should be able to do in context. As one interviewee explains: *"We need to be very explicit about what this autonomy entails. When does the system wait to fire? Where does it go? When does it turn left or right? What do we even want from this autonomy? … We want to have this autonomy under control."* The cultivation of trust, both between soldiers and within the human-machine relationship, is equally foundational. *"If we want to move towards autonomous systems, they have to become part of our DNA, just as our drills and skills are now. And you already mentioned it at the beginning, trust. Trust is incredibly important. Also, between soldiers themselves and from soldiers towards the system."* This shift is not optional. The relentless momentum of technological development and international military trends means that the Dutch armed forces must adapt or risk existential challenges. As another respondent noted, "*Yes, this is simply an extremely big change. And then we go there, because the way I see it, and I don't know how you see it, but I think these technologies are all being developed. So, it's not a question of whether we want to or not, but we have to work with them as the armed forces."* The urgency of this paradigm change is further illustrated by operational realities in ongoing conflicts: *"Now we see, especially in Ukraine, when it comes to UAS and unmanned systems. There is simply continuous innovation. There is a continuous race. Systems are continuously being adjusted, so you no longer buy a*

system for thirty years. That realization has now also emerged within the Defence top and the Defence organization as a whole."

The implications of this transformation are profound. Not only does it require technological innovation, but it also calls for deep changes in procedures, training, and the very definition of military effectiveness. Incremental approaches, characterized by long procurement cycles and static, one-off acquisitions, must be replaced by dynamic, iterative processes of development, testing, and adaptation. This new way of working demands that trust, user input, and organizational learning become embedded within the RNLA's culture, ensuring that autonomous systems are continuously aligned with both operational realities and ethical standards. In sum, the paradigm change from static procurement to continuous, adaptive integration of autonomy is both inevitable and foundational. It serves as the backdrop for the subsequent themes: transformations in legal and ethical review, new approaches to testing and certification, evolving organizational skills and training, and deeper, more flexible collaboration with industry. The following sections explore each of these domains in detail, building on the rationale established here.

The introduction of AWS places significant pressure on the traditional legal review processes within the RNLA, particularly those associated with Article 36. As explored in the previous results, continuous technological innovation and frequent system updates undermine the established pattern of conducting legal reviews as one-off, pre-deployment procedures. One respondent captured this shift: "*Continuous innovation… Systems are continuously being adjusted, so you no longer buy a system for thirty years."* Several participants questioned whether legal procedures developed years ago are still appropriate for short-cycle innovation and constant software updates, with one suggesting: *"We need to scrutinize that procedure and see if it still fits. If a systems' functioning or effect changes significantly, you really do need to do another review."* These observations echo recent literature which highlights the challenge of keeping oversight mechanisms current in rapidly evolving, data-driven military systems (Seixas-Nunes, 2020). There is broad recognition that these challenges are new and only partly addressed by existing procedures. One legal officer reflected: "*If you're talking about the Legal Review... a lot has to change, especially because working with AI is something we've never done before. And when it comes to datasets, AI reliability, and explainability, these are such new things that they need to be formally established. That doesn't exist yet, or only because we've recently developed it."* A further complication is that AWS development is ongoing, even after initial acquisition and deployment. As one interviewee noted: *"But even after it's finished, the AI will continue to develop. There will be software updates. I don't think anyone would ever find a continuous machine learning process, while the system is operating, acceptable. But it could be further developed, at some point, there might be an AI update in the system that allows it to do certain things better, but also differently. If its functioning or effect changes significantly, you really*

do need to do another review." This evolving technical reality has prompted internal debate over how best to organize these ongoing legal reviews. As described*: "The current questions being discussed are: should such a review always go back to the legal review committee, or should you create some sort of forward-deployed posts, where the people who are actually involved with the developments and use of these systems are trained to perform a kind of quick legal review on the spot? Otherwise, you risk slowing down the whole process; it's nice that AI can get fast updates, but if the system then has to sit on the shelf because only one or two people can do a review, that doesn't help. So how this will work in practice is currently under discussion. But I could imagine, for example, that legal advisors who are forward deployed in operational units would also receive legal review training, so that if they become aware of an update, they immediately raise the issue, ask for clarification on what has changed, and check if it's acceptable from a legal perspective."* Additionally, legal reviews are increasingly dependent on the nature and quality of data used to develop and update AWS. "*If you set high requirements for performance in operations, and I think you should, you automatically set high requirements for your data: quality of data and your data process, including setting up TEVV. Otherwise, you are not acting responsibly."* With AI, *"you want to retrain it… if the system gets a software update, it can become a completely different system, and you haven't tested that. That's the risk, AI is easily updated, and you want to update during the lifecycle, as your training set can become outdated. That's something to be aware of. It's not yet perfectly regulated, a real point of attention."* This links directly to earlier findings on the organizational challenge of maintaining oversight of rapidly changing data environments.

In sum, the findings reveal that while the Article 36 legal review procedure remains an essential safeguard, it is increasingly challenged by the continuous evolution and data-driven character of autonomous systems. Both the scope and the process of legal review are under pressure to become more dynamic, context-specific, and iterative, in line with the rapid pace of technological development and operational demands. This organizational adaptation is still in progress and highlights a growing awareness that oversight mechanisms must evolve alongside the systems they govern. The emerging debates about how, when, and by whom legal reviews should be performed also underscore a broader organizational shift: the need for new forms of expertise, collaboration, and procedural flexibility to maintain effective control. This tension between static review procedures and the dynamic reality of autonomous systems directly extends into the operational phase, where commanders, operators, and technical personnel must navigate not only legal requirements but also questions of trust, reliability, and human-machine collaboration.

The commander's ability to responsibly deploy AWS is fundamentally dependent on the assurance that these systems will behave reliably and predictably within the operational context. This assurance is not a

given; it must be actively produced and maintained by the organization through rigorous, ongoing processes of testing, evaluation, verification, and validation (TEVV). As reflected in the interviews, commanders do not simply rely on technical documentation but require confidence that AWS have been subjected to comprehensive, scenario-based testing that mirrors the realities of contemporary conflict. Reliability, in this sense, is inseparable from the quality and representativeness of the training and testing data. As one interviewee noted*: "If you set high requirements for performance in operations, and I think you should, you automatically set high requirements for your data: quality of data and your data process, including setting up TEVV. Otherwise, you are not acting responsibly."* The interview data further highlight that the responsibility for TEVV is not confined to individual engineers, external developers, or a specialized technical team. Rather, as one respondent asserted*, "the responsibility for whether the neural network performs well is test, evaluation, verification, validation responsibility. That is an organizational responsibility."* This distinction marks a fundamental shift in how the RNLA must approach technological oversight. TEVV is no longer a box to be checked at the procurement stage or a purely technical exercise managed by contractors. Instead, the entire organization must be invested in ensuring the integrity, reliability, and trustworthiness of AWS across the system's lifecycle. Several interviewees linked this institutionalization of TEVV directly to the effectiveness and legitimacy of AWS deployment. By embedding TEVV processes across departments, technical, operational, and command, the organization seeks to maintain a unified standard of reliability and accountability. This approach also allows for feedback from operators and commanders in the field to directly inform testing priorities and system improvements, reinforcing a continuous learning cycle. In this way, TEVV becomes an enabler for both operational flexibility and ethical compliance, making it possible for commanders to rely on AWS performance in high-stakes environments. This collective responsibility reflects a broader organizational adaptation: the recognition that oversight of autonomous systems is not merely a technical or contractual matter but a shared institutional commitment. The integrity of AWS, and by extension the trust placed in them by commanders and operators, depends on a transparent, organization-wide approach to TEVV, one that is continually updated and scrutinized in light of real-world experience and operational feedback. By integrating these processes, the organization enables commanders to make informed, context-sensitive decisions about AWS use, supporting both operational effectiveness and ethical accountability.

Empirical findings reveal a recurring theme: effective and trustworthy deployment of AWS hinges on a new blend of knowledge, skills, and relationships, transcending established professional boundaries. A central insight from the interviews is the acute shortage of personnel capable of bridging the gap between military operations and programming expertise. As one interviewee stated, *"You need a specific kind of people. People who know what needs to be programmed from a military perspective, but who can also*

program. So, I need soldiers with programming knowledge, or programmers we can train with military knowledge. They literally have to put in the ones and zeros that determine the behaviour of an autonomous system." This need extends well beyond basic technical competence; it implies cultivating "hybrid" professionals who are able to understand both the requirements of a mission and the inner workings of an algorithm. The current experience-based culture of the RNLA, where *"ranks represent not only hierarchy, but also experience,"* must therefore adapt to accommodate the increasing relevance of technical innovation. As another respondent explained, "*The higher the rank, the more insight you have into the behaviour and context of the underlying levels. … That applies to the tactics and behaviour of autonomous systems too."* Yet, integrating such expertise into the hierarchical structure is not straightforward. The interviews repeatedly highlighted the challenge of bridging language and culture barriers: *"There are really not many people who understand both. … There is a language and culture barrier. A lot of time and effort has to be put into that."* The lack of adequate expertise within the information can provide to jeopardize the ability to crucially scrutinize the data and algorithms for behaviour incongruent with IHL. These findings indicate that fostering an adaptive, learning-oriented culture is as important as technical development. Organizational routines, training programs, and professional development initiatives must be deliberately restructured to bring together legal experts, operators, programmers, and commanders. Continuous professionalization in this domain is essential, especially as legal, technical, and operational uncertainties increasingly overlap. The result is a cultural paradigm shift: from isolated pockets of expertise and experience to multidisciplinary, communicative teams capable of learning together and adapting to fast-changing operational, ethical, and technical landscapes.

The demand for continuous TEVV and rapid innovation also compels the RNLA to radically rethink its relationship with industry. The traditional, transactional acquisition model, where finished systems were purchased and used largely unchanged for decades, has proven untenable amid the pace of technological development in unmanned and autonomous systems. As one senior officer explained, "*Now we see, especially in Ukraine, when it comes to UAS and unmanned systems, there is continuous innovation. There is a continuous race. Systems are continuously being adjusted, so you no longer buy a system for thirty years. That realization has now also emerged within the Defence top and the Defence organization as a whole. … Ideas are being formulated that procurement and acquisition must be organized differently. … We must be able to procure more quickly. We need to develop more together with those companies. We need to build a more direct relationship with companies, working with, for example, a drone ecosystem."* This "drone ecosystem" model exemplifies the shift to direct, dynamic, and co-developmental relationships with commercial partners. The operational context now requires the military to collaborate actively with industry, not only during initial system development but through continuous

cycles of iteration and adaptation. As one interviewee confirmed, "*You see more direction towards closer collaboration during the development process? Yes, during the development process. But that development process takes place continuously. It does not stop. … If a thousand drones have been developed, it does not stop there. Because we just keep developing afterwards."* However, such collaboration introduces significant new challenges. Data ownership, intellectual property rights, and the transparency of proprietary algorithms become critical, as do questions about aligning civilian technological innovation with military requirements. One respondent highlighted this complexity: "*These are all those challenges that need to be looked at. The same applies to who owns the intellectual property of what is developed. … There is an enormous overlap in ethical, legal, and economic interests. This is totally new. … It is a paradigm change, is really totally different."* The interviews also suggest that, up until recently, the military has adopted a largely passive role, adapting to whatever industry supplied: "*It's been more industry tells us what they want, and then we just do it. We adapt ourselves to the technology."* To meet these new demands, the RNLA must become a far more articulate and proactive partner. This involves specifying operational requirements with greater clarity, demanding transparency wherever possible, and cultivating the organizational expertise necessary to critically assess and influence technical development. As one interviewee described, *"Part of what I'm trying to do is get the military to be able to more eloquently define what it is they want and come to the table with industry and be able to say, this is what we need. We need you to keep the data open. We need to own the data that you're developing on. Really become a leading force in that discussion, partly because we are warfighting experts, and they are the tech experts."* Moreover, the fragmented nature of the Dutch defence industry and dependence on international suppliers, compounded by complex European procurement rules, make building and sustaining such partnerships even more challenging: "*This is the Netherlands. We have little defence industry in the Netherlands. … The military industry we have is commercial, is civilian. And therefore, it is fragmented. … We have always chosen to develop as little as possible ourselves. And just go along with the market. With suppliers, with the Americans, with the Germans. … If you want to know what that means, those competition clauses."*

However, this shift toward continuous adaptation and co-development with industry directly challenges the historically centralized procurement structure within the RNLA, and subsequently the whole Dutch military organization. For decades, procurement has been governed by a clear division of labour: the COMMIT agency handles all direct relations with manufacturers and suppliers, acting as the sole point of contact and negotiation for equipment acquisition. As described in one interview: "*Now, within Defence, there is a fairly strict separation. So, as part of Defence, you have COMMIT. That is actually one of the companies within Defence and they are responsible for purchasing, acquiring equipment. This means that they have direct contact with manufacturers. They conduct the negotiations, can make commitments. And*

the agreement is that, actually, only COMMIT does this, and the rest of the organization does not make direct agreements with the industry." While this model was effective for long-term, hardware-based acquisitions, its suitability for supporting integrated and continuous TEVV processes is increasingly in question. The new paradigm demands not only rapid procurement, but also ongoing collaboration and feedback loops between operational users, developers, and legal/technical experts throughout the lifecycle of each system. Integrated TEVV processes require that lessons from real-world deployment and end-user experiences directly inform the design, certification, and evolution of AWS, often in real time. When a central agency like COMMIT intermediates all industry contact, it risks creating silos, slowing down essential feedback, and hampering the direct, iterative communication that is vital for adaptive system development Procurement, operational, legal, and technical domains must work together continuously, blurring the boundaries that a strictly centralized model was designed to enforce. As the empirical evidence shows, the shift to continuous, user-driven adaptation "*must be organized differently,"* and this may require fundamentally rethinking the structure and role of procurement agencies like COMMIT within the broader organizational ecosystem.

In sum, the increasing integration of TEVV and the need for dynamic, co-developmental industry partnerships challenge the traditional separation of procurement from operational and technical users. For the RNLA to keep pace with the speed of technological innovation and ensure user-centred, ethically responsible autonomy, the organization may need to move beyond a model in which a central institute like COMMIT controls all industry engagement.

Finally, we investigate what organizational changes are needed in the accountability cycle. As established in Theme 1, the RNLA's accountability architecture, anchored in AARs, BDAs, and oversight by judicial authorities such as the KMAR and Public Prosecution Service, is designed to ensure traceability, transparency, and legal compliance in the use of force. Yet, the implementation of AWS fundamentally alters both the content and the process of these accountability mechanisms. The shift from human-centred to machine-executed operations requires a thorough re-examination of how data is generated, analysed, and used for retrospective evaluation of military actions. Traditionally, AARs and BDAs have relied heavily on human recollection, immediate reports, and subsequent physical or digital investigation. However, as several interviewees noted, human memory is fallible especially under pressure or when legal consequences are at stake: *"We have a very selective memory. That's proven. Especially when there is possible punishment in play. We are notoriously bad at reliably reconstructing events from the past"* This limitation becomes critical in high-stakes incidents, where the completeness and accuracy of post-action reporting are essential for both internal learning and legal accountability. By contrast, *"an autonomous system by definition needs a camera. So that data is there. You can log it. And if something*

has gone wrong, or right, you can improve the system, because you have the data". This means every decision point and sensory input can, in principle, be recorded and extracted: *"If all goes well, the system comes back from the mission and then you can pull those logs out".* As a result, system logs and sensor data offer a much more granular and objective account of what happened during a mission, supplanting the subjectivity and incompleteness of human testimony. *"During execution, everything can be recorded. And that is really a big difference from humans".* This new standard of evidence presents both opportunities and challenges. On the one hand, logs and recordings allow for more precise reconstructions of events: "*With those messages you can replay the action as best as possible in the simulation, including the context... You open the box; you get the log data out. You can replay the mission exactly within the boundaries of the simulation. There's no bias or reticence. In that sense, it's much more transparent"*. On the other hand, retrieving this data is not always straightforward. For AWS that are destroyed in combat such as loitering munitions it is less possible to extract information: "*Then it didn't complete its mission. So, you already know something went wrong. But yes, that's the same as with a human. If a soldier doesn't return, you know the mission failed".* For truly autonomous, self-destructive systems, data loss is a significant problem. "*How do you get those images back? Maybe you can't, unless you have another sensor externally measuring what happens. But then you've essentially made an uncontrollable system. And the question is whether you would even want to have that"*. As a solution, respondents suggested the need for accompanying observation drones or external surveillance assets to ensure that mission data is captured even if the AWS itself is lost: *"That's why it's useful to monitor with a drone. So that's good organization".* Beyond mere retrieval, AWS logs have the potential to streamline and even automate parts of the reporting process: *"You want, when this has been used, for the data behind it to be stored. And that at the end of the day, with the press of a button, you see an overview of these targets I've indicated at this time, at these coordinates, and that it's sent with one button to his higher command, who uses it in the larger reporting. So those are things that are what I call low-hanging fruit. My estimate is these are relatively easy things you could set up"*. In principle, AWS can be configured to automatically produce standardized reports, significantly reducing the time and labour involved in traditional AARs and BDAs, and minimizing human error or intentional omission. Yet, even as these systems promise greater accuracy, they also require new forms of technical and legal expertise. Investigating incidents involving AWS demands the ability to interpret system logs, algorithms, and machine learning outputs. As one respondent put it: *"Well, on the legal side you need people who can read technical information. All the tooling they have. Like a simulation to replay the action. Because with those messages you can replay the action as best as possible in the simulation, including the context"*. The integration of AWS thus calls for a new multidisciplinary approach in accountability investigations. Legal and military investigators must now include technical experts capable of extracting, analysing, and contextualizing digital evidence: *"You*

need people who understand AI. Who has an overview. Who understand what the risks are and when those systems make mistakes. And what information they need. And under which conditions they don't work well". These experts are essential not only for post-mission analysis, but also for real-time operational safety and risk management. Furthermore, the possibility of using simulation environments to re-play missions offers a new standard for forensic investigation: *"With those messages you can replay the action as best as possible in the simulation, including the context. That's the beauty. I think that is also a fundamental difference compared to humans"*

Thus, the integration of AWS into the RNLA's operations necessitates a profound transformation of the existing accountability process. While the architecture described in Theme 1, centred on AARs, BDAs, and judicial oversight, remains fundamentally relevant, its implementation must evolve to address the technical realities and unique risks of autonomous systems. The shift from subjective, human-centred reporting to objective, data-driven accountability presents both opportunities for greater transparency and challenges in terms of data retrieval, technical expertise, and responsibility allocation. Yet, as highlighted throughout this research, it remains of paramount importance that there is a clear, causal track of human control, a transparent line connecting decision-making, system configuration, and the operational actions taken by AWS. This is not only vital for upholding legal standards and enforcing responsibility, but also for ensuring public trust in military operations involving autonomy. Without the ability to trace outcomes back through deliberate human choices, meaningful accountability risks being lost in technical complexity. As the Army moves forward, the alignment of accountability mechanisms with the operational realities of AWS will be crucial for sustaining both ethical and legal legitimacy in future conflict.

5 Discussion

This research examined the feasibility of applying the CHOF, supplemented by the Glassbox Framework, to ensure ethical and accountable deployment of LAWS within the RNLA. Through systematic analysis of theoretical predictions against empirical evidence from RNLA practices, this investigation reveals a complex landscape where oversight frameworks face both confirmatory challenges and unexpected nuances in their application to autonomous weapons systems. The discussion proceeds through the three sub-questions that structured this research, each building upon the previous to construct a comprehensive assessment of current oversight capabilities and future possibilities. Together, these analyses point toward a differentiated approach to AWS oversight, one that recognizes varying degrees of ethical tractability across autonomous functions while identifying specific technical and institutional prerequisites for meaningful human control. The following discussion examines each sub-question's findings systematically, revealing how empirical evidence both confirms and refines theoretical frameworks, and ultimately charting pathways for implementation that neither reject nor uncritically embrace autonomous weapons technology.

5.1 Findings

The first sub-question examined what principal ethical challenges are posed by LAWS and how these are reflected in current RNLA oversight practices. The literature converges on the claim that LAWS, particularly those using machine learning, strain core jus in bello principles of distinction, proportionality, necessity, and humanity because they lack contextual moral judgement and instead rely on correlational inputs, which are vulnerable in dynamic or degraded environments (Boutin & Woodcock, 2024; Hughes, 2020; ICRC, 2019; Kwik, 2024; Schwarz, 2021). These technical limits erode the quality of human judgement at the point of action and risk translating normative assessments, such as “excessive” harm, into quantifiable proxies with diminished ethical meaning. A second, closely related prediction is the emergence of an accountability vacuum. Traditional chains of command presuppose identifiable human agents and constructible decisions. Autonomy and opacity disperse causal links across developers, operators, commanders, and algorithms, while black-box models undercut traceability and explainability, both preconditions for legal and administrative forums to assess wrongdoing and attribute responsibility (Bovens, 2007; Cavalcante Siebert et al., 2023; Christie et al., 2024; Santoni de Sio & Van den Hoven, 2018). Third, the literature holds that Article 36 weapons reviews are ill-suited to adaptive or online-learning systems: one-time pre-deployment checks cannot anticipate distribution shift or post-deployment drift, implying a need for continuous reassessment and auditable, output-based evidence (Christie et al., 2024; Poitras, 2018; Verbruggen & Boulanin, 2017b). If these risks manifest in practice, we should

observe stress on targeting cycles, legal reviews, explainability, logging, and post-strike inquiry. The empirical findings broadly confirm these theoretical concerns. Within the RNLA, Article 36 reviews remain the doctrinal cornerstone of ethical oversight, but they are not yet adapted to the characteristics of adaptive AI systems. The explainability and opaqueness of ML systems limit legal reviewers' ability to pre-emptively assess operational behaviour, as performance depends on data quality, domain alignment, and contextual fit. No formal triggers exist for re-review when models are updated, data inputs change, or systems operate outside their intended design parameters. TEVV practices for AI-enabled systems are uneven, with gaps in drift detection, uncertainty calibration, and data provenance. The targeting cycle, the RNLA's operational process for translating IHL principles into real-time decision criteria, is particularly vulnerable, as it relies on human moral reasoning, contextual awareness, and experiential judgement. AWS cannot ethically deliberate, interpret ambiguous or novel situations, or weigh alternative courses of action. Technical limitations, such as a lack of causal reasoning, unpredictable behaviour in novel contexts, and the black-box nature of decision-making, further undermine both traceability and explainability. While formal rules of engagement preserve human authority to intervene, empirical evidence shows this authority can be nominal in time-sensitive or communications-degraded contexts. Without interfaces that surface uncertainty and system limits, human operators risk becoming passive authorisers rather than active decision-makers. The accountability process is equally strained by LAWS integration. Agency becomes distributed across developers, integrators, vendors, operators, and commanders, and due to lack of explainability and traceability, it is impossible to isolate a single responsible agent. When commanders deploy LAWS, they risk being held accountable for mis usages that do not stem from their direct responsibility and state only to use LAWS when found reliable. The retributive logic of criminal justice is ill-suited to non-human actors, and sanctions against corporate entities cannot substitute for individual culpability. Enforceability is further weakened by the opacity and unpredictability of AWS behaviour under novel conditions, which challenges the foreseeability and control needed to attribute fault. Paradoxically, AWS can generate detailed telemetry, sensor logs, and configuration states that, if preserved, standardised, and interpretable, could improve post-action investigations and even automate parts of the AAR/BDA process. However, data loss (especially with expendable munitions), proprietary restrictions, and a shortage of multidisciplinary investigative capacity currently limit these benefits. The findings substantiate the literature's core predictions: LAWS threaten IHL compliance by weakening the human moral agency embedded in targeting cycles and by dispersing accountability in ways existing legal frameworks are ill-equipped to handle. Both theory and practice converge on a central condition: commander accountability is sustainable only where foreseeability, reliability, and traceability are demonstrably present. In the absence of these prerequisites, responsibility

risks becoming diffused across the institution or the state, thereby undermining the legal and ethical foundations upon which operational oversight is based.

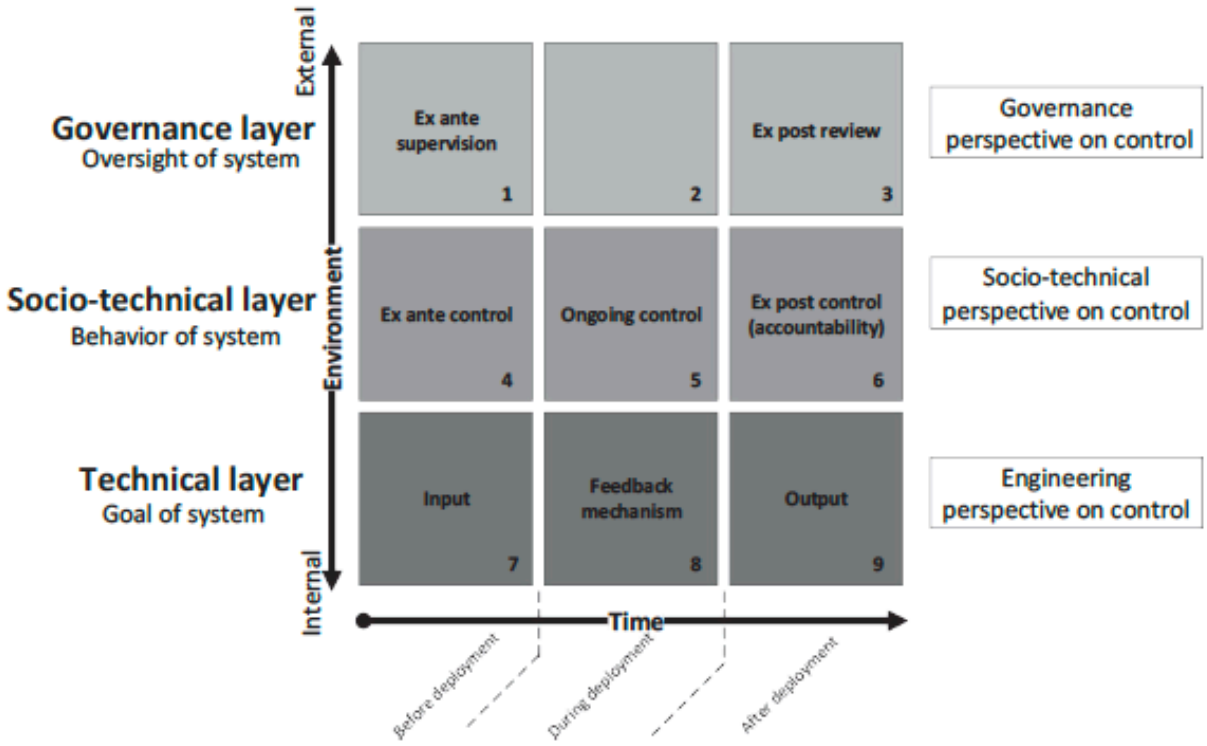

Figure 3. The Comprehensive Human Oversight Framework (CHOF) as developed by Verdiesen (2024), organizing oversight across time (before, during, after) and system environment (technical, socio-technical, governance)

The second sub-question sought to map these identified ethical problems onto the structure of the CHOF framework. For reasons of readability, the CHOF matrix is reproduced in figure 3 to guide the mapping of empirical findings onto its structure. Building on sub-question 1, which revealed how LAWS strain oversight mechanisms such as Article 36 reviews, the targeting cycle, and post-engagement accountability, the purpose of sub-question 2 is to locate these problems within the structure of the CHOF. By mapping observed challenges onto CHOF's three phases and its three control layers, the framework helps to assess gaps in control. The literature highlights two persistent weaknesses during the use phase: the absence of independent, real-time oversight (Block 2) and the lack of ongoing, actionable behavioural control (Block 5). This discussion examines how far RNLA practices align with CHOF's expectations and where the tensions observed in sub-question 1 materialise within its matrix. In CHOF, the pre-deployment phase is designed to produce bounded, evidenced delegation under conditions of executive autonomy.

Governance (Block 1) authorises use through mandate ROE, and legal review, and is expected to tie such authorisation to foreseeable system behaviour. Socio-technical preparations (Block 4) should train people and procedures to enact these bounds, while the technical layer (Block 7) translates ROE and IHL into verifiable requirements, validated performance and defensible thresholds that make authorisation reviewable downstream. The empirical results show that stating technical requirements (Block 7) is possible in principle but fragile in practice, and that fragility propagates into both governance (Block 1) and operator readiness (Block 4). Interviewees emphasise that LAWS performance depends heavily on domain alignment and data quality, which are difficult to guarantee ex ante. Models trained on specific environments may fail in unfamiliar operational contexts, and ensuring congruence between training, testing, and deployment conditions is costly and only ever partial. This makes thresholds policy-laden and hard to calibrate. Stakeholders recognise the appeal of rules such as "only engage if confidence > x," but question what any score means for actual misidentification risk. Even well-trained models can fail under domain shift, and there are no accepted metrics to relate, for example, a 0.95 confidence score to the tactical quality of a human identification decision. The opacity of machine learning compounds this uncertainty, as the results show that engineers are unable to show how LAWS use data to make decisions. Thresholds can be set, but their meaning is contested, and they cannot be normatively validated against IHL ex ante. These technical limits have direct implications for governance and socio-technical preparation. Legal reviewers under Article 36 may define the ethical and legal values LAWS must adhere to, and commanders can be trained on ROE and doctrinal requirements, but ultimate adherence depends on how reliably the system itself can operationalize these norms when authority is delegated. Without reliability, predictability and stable interpretability, Block 1 cannot ensure bounded authorisation, and Block 4 cannot train operators on cues that map onto system behaviour. In short, pre-deployment governance remains strong in form but fragile in substance, pushing much of the oversight burden upstream without securing evidentiary assurance.

The accumulation of these problems becomes more visible in the deployment phase. CHOF distinguishes three layers here: Block 8 (technical), where the AWS runs feedback loops against pre-set goals and constraints; Block 5 (socio-technical), where humans set goals but do not determine concrete actions during execution; and Block 2 (governance), where an independent mechanism should monitor actions in real time. In theory, this triad ensures that autonomous functions remain constrained by pre-set norms while retaining human and institutional oversight. The empirical findings reveal that, in line with theory, blocks 2 and 5 are comprised of effective human control. Because of the autonomous character of LAWS, there is no mechanism for ongoing control (Block 5) and no means for independent, real-time oversight (Block 2). Doctrinally, the targeting cycle already embeds IHL principles, target development, validation, proportionality, execution, and assessment, and respondents describe it as a robust tool for legal

compliance. Yet once targeting and engagement decisions are delegated to autonomous functions, governance during deployment becomes derivative of what was authorised in pre-deployment. In practice, this means that commanders should rely on the LAWS to operate exactly according to how requirements are set beforehand. The post-deployment phase is meant to convert battlefield outcomes into accountability and institutional learning. Governance (Block 3) should reconstruct what happened, allocate responsibility, and impose remedies; the socio-technical layer (Block 6) should translate findings into doctrine, training, and procedures; and the technical layer (Block 9) must furnish auditable artefacts, logs, model states, datasets, that make the review evidential rather than declarative. The RNLA possesses mature review practices in form, such as After-Action Reviews, Battle Damage Assessments, and legal follow-up procedures. Yet algorithm opaqueness and data ownership constraints weaken the evidentiary foundation these forums require. Without reliable and predictable insights on algorithm functioning, reconstructions risk becoming partial or speculative. Updates and retraining exacerbate this challenge, as the system reviewed under Article 36 may differ materially from the one deployed. Respondents acknowledge this fragility and suggest compensatory practices, such as observation drones to record engagements, but these measures are piecemeal. The result is a post-deployment phase that is procedurally robust yet substantively fragile. Forums for accountability exist, but the inputs they need are not consistently available or interpretable. Autonomous systems can, in principle, generate time-stamped logs and automated reporting, but many platforms (e.g., loitering munitions) are expendable, raising the risk of data loss. Without reliable technical artefacts, governance and socio-technical review risk becoming symbolic, offering process without substance. The results therefore refine the literature's critique: while gaps in Block 2 and Block 5 are confirmed, the findings also expose a critical vulnerability in Block 7. Reliable governance requires that LAWS not only be tested but also remain traceable and trackable across contexts (Santoni de Sio & Van den Hoven, 2018). Trackability is undermined because the fragile technical requirements prevent systems from reliably aligning with contextual environments, meaning commanders cannot reliably make pre-deployment adjustments. Oversight is reduced to static thresholds that are easily invalidated by data inadequacy. Traceability is compromised because unstable technical requirements break the causal chain needed to follow a system's decisions from Article 36 review, through battlefield use, to post-deployment accountability forums. Together, these weaknesses reveal that CHOF's promise of bounded, evidenced delegation rests on fragile technical ground: unless traceability and trackability are secured, the framework risks offering oversight in form but not in substance.

The third sub-question evaluated to what extent the integration of the Glassbox Framework with CHOF can address the identified ethical and operational gaps in LAWS oversight. The Glassbox Framework strengthens CHOF by reorienting oversight away from inaccessible internal model logic and toward

observable input–output behaviour. Its central premise is that accountability can be secured without full explainability by focusing on what a system does, rather than how it computes internally. To this end, Glassbox structures oversight in two stages: interpretation, where abstract principles such as proportionality or rules of engagement are translated into measurable behavioural requirements; and observation, where compliance is monitored through system outputs, telemetry, and logs. This approach is attractive in military settings like the RNLA, where decision cycles are compressed, technical expertise is unevenly distributed, and external validation of machine learning models is often infeasible. By anchoring oversight in observable behaviour, Glassbox provides non-technical authorities with accessible mechanisms for identifying violations and triggering escalation. The results show that some functions are indeed less ethically challenged, and one could argue that these norms can be specified and verified through input–output monitoring alone because they concern externally legible behaviours rather than the system's internal reasoning. Examples include functions that are related to the navigation, ISR or even some functions within swarming domains, such as geofencing and no-strike areas, temporal spacing between engagements, arming states, abort-on-uncertainty rules, and logging completeness. For these, oversight can compare observed outputs (e.g., no fires outside polygon X; engagement only after human confirm; abort when sensor fusion confidence drops below Y) to pre-specified requirements and reach defensible compliance judgments without opening the model. By contrast, targeting norms, notably distinction and proportionality assessments which hinge on the quality of target identification, are constituted by technical requirements that presuppose reliability and predictability. Here, the normative claim ("engage only combatants") reduces, in practice, to properties of the sensing-and-classification pipeline: domain alignment, training/test data, calibrated scores whose numerical values correspond to real misidentification risk, validated performance within a defined operational design domain, robustness under degradation (EW interference, clutter, adversary deception) and stable behaviour across updates. These conditions cannot be inferred from outputs alone unless prior TEVV work has established that the observed outputs are trustworthy indicators of compliant internal functioning. Removing that evidential base means input–output traces can flag violations after the fact but cannot anchor ex ante authorisation or ethically meaningful thresholds for lethal action. In short, input–output monitoring can derive and enforce procedural/behavioural norms, but targeting norms are epistemic: they depend on guarantees about how the system forms its classifications. Meeting them requires demonstrated reliability and predictability, not merely observable compliance. Framed through Santoni de Sio and Van den Hoven's (2018) account of meaningful human control, the argument can be sharpened as follows. Glassbox commendably operationalises CHOF where norms are procedural/behavioural and externally legible: by translating rules into testable requirements and auditing outputs, it helps ensure that system behaviour tracks the relevant human reasons (e.g., geofencing, arming states, abort-on-uncertainty) and supports tracing to identifiable

human decision makers. In these domains, input–output monitoring supplies sufficient evidence that the system is responding to the morally salient features of the situation as intended, thus satisfying both conditions for meaningful control. By contrast, targeting norms (distinction, proportionality) are epistemic: compliance depends on whether the sensing–classification pipeline reliably and predictably covaries with the morally relevant facts (domain alignment, calibrated and validated scores, robustness under degradation, stability across updates). On Santoni de Sio and Van den Hoven's view, such reliability is a precondition for genuine tracking; remove that evidential base and observable outputs risk reflecting spurious correlations, and Glassbox cannot anchor ex-ante authorisation or ethically meaningful thresholds for lethal action. Moreover, without demonstrable reliability, tracing becomes fragile: commanders cannot plausibly "own" outcomes they are not in a position to understand or endorse. Christie et al. (2024) sharpen this critique by distinguishing formal traceability (what the system did) from substantive traceability (why it acted that way): input–output monitoring may expose violations ex post but does not furnish the epistemic basis commanders require to authorise lethal action ex ante. Consequently, Glassbox can plausibly reinforce CHOF in Blocks 2 and 5 for procedural control, but it cannot, on its own, remedy Block 7's technical fragilities that underwrite targeting; satisfying tracking and tracing there requires opening the black box through TEVV that establishes reliability and assigns accountable human ownership.

Taken together, the findings indicate that the Comprehensive Human Oversight Framework, supplemented by the Glassbox Framework, can only partially ensure ethical, accountable, and operationally effective deployment of LAWS in the RNLA. On the one hand, the frameworks are applicable as diagnostic and structuring tools: CHOF provides a systematic architecture that highlights oversight gaps across governance, socio-technical, and technical layers, while Glassbox operationalises oversight for functions whose normative content is procedurally legible. In domains such as navigation and ISR, the combination of CHOF and Glassbox enables oversight based on observable outputs and auditable behavioural requirements. On the other hand, the frameworks remain insufficient when applied to targeting norms such as distinction, proportionality, and necessity. These are not purely behavioural but epistemic, and therefore presuppose demonstrable reliability, predictability, and traceability of the underlying classification pipelines. Translating such normative thresholds into enforceable technological requirements is only credible if algorithms are thoroughly tested, validated, and stabilised under conditions of operational relevance. This necessitates opening the black box: without transparent access to system behaviour neither CHOF's bounded delegation nor Glassbox's input–output monitoring can substantively guarantee ethical compliance. Thus, the extent of applicability is conditional: the frameworks are practically valuable for procedural norms but remain fragile for lethal targeting functions. This overarching gap explains why the RNLA encounters difficulties in conducting a credible ex ante

Article 36 review, why targeting decisions during deployment remain vulnerable to reliability and proportionality failures, and why post-deployment accountability cycles risk breaking down in the absence of traceable, substantive evidence.

5.2 Practical Recommendations

To operationalise the oversight envisioned by CHOF and supplemented by Glassbox, the RNLA must enact a transformation across acquisition, testing, and accountability practices that directly addresses the technical fragilities exposed in this study. The current model of linear procurement and one-off legal reviews is inadequate in the face of adaptive, machine-learning–enabled weapon systems. Instead, an organisational paradigm of continuous, scenario-realistic TEVV is needed, where reliability and predictability are institutionally produced rather than presumed. This requires not only iterative testing and validation but also structured triggers for re-review when models, data, or operational design domains change. Central to this transformation is the recognition that proportionality and necessity cannot be reduced to static thresholds; translating these normative principles into enforceable technological requirements demands opening the black box through algorithmic transparency, validated calibration, and traceable data lineage. To support this, the RNLA must develop hybrid expertise that bridges technical and operational domains, ensuring that commanders are not reduced to passive authorisers but remain capable of exercising meaningful control. Data governance is equally critical: ownership of datasets, model states, and telemetry must rest with the RNLA to guarantee both operational sovereignty and forensic accountability. This also implies redefining procurement partnerships, moving beyond transactional acquisition toward collaborative co-development that secures transparency over algorithms and system updates. Finally, accountability mechanisms must evolve to integrate standardised and accessible system logs into AAR and BDA processes, strengthening the evidentiary base for legal and ethical review. Only by embedding these adaptations, continuous TEVV, black box opening, hybrid expertise, robust data ownership, and collaborative procurement, can the RNLA transform CHOF and Glassbox from frameworks that are formally robust but substantively fragile into a credible foundation for ethical, accountable, and operationally effective deployment of autonomous weapon systems.

5.3 Limitations

This study is subject to several limitations, which are acknowledged to promote transparency and contextualize the scope of its findings. These limitations stem primarily from the study's research design, the status of the technology under investigation, access constraints, and the interpretive nature of the analysis. First, as a single-case study, the research focuses exclusively on the RNLA and does not aim for broader generalization. While the findings may offer analytical insights applicable to other military or institutional contexts, their transferability depends on the degree of contextual similarity. This is a known

trade-off in in-depth qualitative research and is addressed through thick description and methodological transparency. Second, the autonomous weapon technologies examined are not yet fully operational within the RNLA and currently possess a relatively low Technological Readiness Level (TRL). Consequently, much of the data reflects anticipatory governance, policy frameworks, strategic planning, and normative expectations, rather than lived institutional practices. This limits the study's ability to evaluate the actual implementation and efficacy of oversight mechanisms in deployed systems. Third, although most information obtained through interviews and document analysis is not classified, it is often the classified dimensions of system architecture, operational procedures, or internal reviews that are most critical for assessing ethical oversight in practice. The inability to access such sensitive material restricts the study's capacity to fully trace how accountability and human control are enacted in real-world scenarios, particularly within the technical and sociotechnical layers of the CHOF framework. Fourth, the use of semi-structured interviews introduces potential issues related to self-representation and selection bias. Interviewees may emphasize particular narratives aligned with institutional priorities or personal viewpoints. This is mitigated through source triangulation and a critical interpretive approach during analysis. Finally, as an interpretivist, qualitative inquiry, the research is inherently shaped by the positionality of the researcher. While reflexivity is practiced throughout, the possibility of interpretive bias remains. This is addressed by maintaining transparency in analytic procedures, using traceable coding practices, and grounding findings in a clearly articulated theoretical framework. At the same time, these methodological limitations should be read in light of the researcher's positionality and the ethical commitments shaping coding and interpretation. As a researcher, I occupy a dual position: external to the RNLA yet working in close dialogue with military stakeholders. This position required balancing critical distance with sensitivity to institutional realities, a tension that inevitably influenced both the framing of research questions and the interpretation of interview material. The coding process, while guided by established procedures of thematic analysis, was therefore not neutral. Decisions about which excerpts to prioritize, how to cluster themes, and how to link them to theoretical constructs were shaped by my interpretive lens and normative commitments to accountability, human oversight, and compliance with international humanitarian law. Recognizing this subjectivity is essential: the results do not claim to represent an objective truth about AWS oversight, but rather a situated, reflexive diagnosis of how CHOF and Glassbox resonate with and challenge current RNLA practices.

5.4 Recommendation for Future Research

The findings of this study reveal persistent gaps and emerging complexities in the ethical oversight of AWS, highlighting not only areas for future research but also the structural conditions that must be met if CHOF and Glassbox are to function as more than formal frameworks. Central among these is the problem

of data domain alignment: the degree to which training, validation, and operational data align with real-world conflict environments. As shown in this study, even rigorously tested systems become unpredictable and ethically fragile when deployed outside their intended design domain. Addressing this requires systematic approaches to scenario-based TEVV, feedback loops that translate field experience into data updates, and ongoing validation of model states across operational contexts. Yet data congruency alone is insufficient. To translate normative thresholds such as proportionality and necessity into enforceable technological requirements, militaries must also open the black box: ensuring algorithmic transparency, calibrated reliability, and traceable update histories that make system behaviour explainable and attributable. This underscores the urgency of developing robust military data strategies that guarantee ownership of critical datasets and models, preventing oversight from being undermined by vendor opacity or proprietary restrictions. Only by aligning technical data practices with institutional accountability structures can the RNLA ensure that CHOF and Glassbox oversight is substantively grounded in evidence rather than symbolic procedure.

Therefore, there is a pressing need for empirical research on the organizational and cultural transformations required to make the oversight structures of CHOF and Glassbox substantively effective in military practice. A central finding of this study is that technological assurance alone is insufficient: sustainable and responsible deployment of AWS demands institutional adaptation in norms, values, and routines. Future research should therefore examine how military cultures, often shaped by hierarchy and rigid chains of command, can evolve toward multidisciplinary collaboration and iterative learning that embed technical expertise into operational decision-making. Central to this is the cultivation of trust, both between human operators and AWS, and across commanders, technical specialists, and legal advisors. As shown here, trust in autonomous systems cannot be assumed but must be actively built through demonstrable reliability, transparent performance, and commanders' ability to understand, predict, and where required, intervene in system behaviour. Because oversight in Blocks 2 and 5 depends on commanders not becoming passive authorisers, future research should investigate how training, professional development, and interface design can empower them to retain meaningful human control. Research should explore organizational practices for systematically incorporating user feedback into AWS adaptation, ensuring that oversight frameworks are not only formally present but also experienced as credible and usable by those who bear responsibility for their deployment.

Thirdly, as accountability mechanisms increasingly rely on automated reporting, digital forensics, and mission data logs generated by AWS, there is a need for research that examines how these forms of evidence interact with the oversight structures of CHOF and Glassbox. While the shift from subjective, human-centred reporting to objective, data-driven documentation offers opportunities for greater

transparency, it also exposes vulnerabilities in data integrity, access, and interpretation. As this study shows, the evidentiary foundation of CHOF's post-deployment phase (Block 9) is fragile: logs may be lost with expendable munitions, constrained by proprietary vendor systems, or rendered opaque by unstable model updates. Future research should therefore explore how military organizations can guarantee the reliability, completeness, and security of mission data across the AWS lifecycle, and how gaps in data continuity can be mitigated through redundancy, contractual safeguards, or technical design. Moreover, the growing use of simulation-based reconstructions raises critical questions about evidentiary standards, procedural safeguards, and the degree to which algorithmic decisions can be meaningfully interpreted by oversight bodies. To prevent accountability from collapsing into symbolic process, further study is needed on how technical experts, commanders, and legal advisors can collaboratively interpret digital evidence, translate algorithmic outputs into accessible forms, and ensure that oversight forums retain both transparency and normative legitimacy.

5.5 Conclusion

This study has shown that while CHOF and Glassbox provide valuable tools for diagnosing oversight gaps in the deployment of autonomous weapon systems, their effectiveness in practice remains limited without deeper institutional adaptation. Ensuring ethical and lawful use requires continuous, scenario-based TEVV, robust data governance, and algorithmic transparency, rather than reliance on static procedures or one-time reviews. Crucially, translating principles such as proportionality and necessity into enforceable technological requirements demands opening the black box, so that reliability, predictability, and traceability are demonstrably secured. Achieving this depends on hybrid expertise, military ownership of data and models, and procurement practices that prioritise transparency and collaborative development. At the same time, military culture must evolve to embed multidisciplinary collaboration, user-centred system design, and trust-building across commanders, operators, legal advisors, and technical experts. As accountability becomes increasingly data-driven, new approaches will be needed to safeguard and interpret digital evidence in both military and judicial forums. Ultimately, ethical oversight in autonomous warfare is not a fixed framework but a dynamic, collective endeavour, one that requires both technical reliability and cultural adaptation if militaries are to uphold their responsibility in the age of machine autonomy.

6 References

AIV, & CAVV. (2021). *Autonome wapensystemen: Het belang van reguleren en investeren*.

Alharahsheh, H. H., & Pius, A. (2020). A review of key paradigms: Positivism VS interpretivism. *Global academic journal of humanities and social sciences*, *2*(3), 39-43.

Altmann, J., & Sauer, F. (2017). Autonomous weapon systems and strategic stability. *Survival*, *59*(5), 117-142.

Amoroso, D., & Tamburrini, G. (2021). Toward a normative model of meaningful human control over weapons systems. *Ethics & International Affairs*, *35*(2), 245-272.

Asaro, P. (2012). On banning autonomous weapon systems: human rights, automation, and the dehumanization of lethal decision-making. *International review of the Red Cross*, *94*(886), 687-709.

Blakcori, N., Stathakis, L. I., Koutsoukos, L. D., & Kirilov, L. K. (2024). The Evolving UAS Threat: Lessons from the Russian-Ukrainian War Since 2022 on Future Air Defence Challenges and Requirements. *NATO, Integrated Air and Missile Defence Center of Excellence*.

Blanchard, A., & Taddeo, M. (2024). Autonomous weapon systems and jus ad bellum. *Ai & Society*, *39*(2), 705-711.

Blauth, T. F. (2023). Autonomous Weapons Systems in warfare: is Meaningful Human Control enough? In *Handbook on the Politics and Governance of Big Data and Artificial Intelligence* (pp. 476-503). Edward Elgar Publishing.

Bode, I., & Watts, T. F. A. (2023). Loitering munitions and unpredictability: Autonomy in weapon systems and challenges to human control.

Bondar, K. (2025). *Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare*.

Booker, J. O. (2024). War Machines of Tomorrow: Accountability and Oversight in the Age of Lethal Autonomous Weapon Systems.

Boulanin, V., & Lewis, D. A. (2023). Responsible reliance concerning development and use of AI in the military domain. *Ethics and Information Technology*, *25*(1), 8.

Boutin, B., & Woodcock, T. (2024). Aspects of realizing (meaningful) human control: a legal perspective. In *Research Handbook on Warfare and Artificial Intelligence* (pp. 179-196). Edward Elgar Publishing.

Bovens, M. (2007). Analysing and assessing accountability: A conceptual framework 1. *European law journal*, *13*(4), 447-468.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative research in psychology*, *3*(2), 77-101.

Castelfranchi, C., & Falcone, R. (2003). From automaticity to autonomy: the frontier of artificial agents. *Agent autonomy*, 103-136.

Cavalcante Siebert, L., Lupetti, M. L., Aizenberg, E., Beckers, N., Zgonnikov, A., Veluwenkamp, H., Abbink, D., Giaccardi, E., Houben, G.-J., & Jonker, C. M. (2023). Meaningful human control: actionable properties for AI system development. *AI and Ethics*, *3*(1), 241-255.

Chengeta, T. (2016). Accountability gap: Autonomous weapon systems and modes of responsibility in international law. *Denv. J. Int'l L. & Pol'y*, *45*, 1.

Christie, E. H., Ertan, A., Adomaitis, L., & Klaus, M. (2024). Regulating lethal autonomous weapon systems: exploring the challenges of explainability and traceability. *AI and Ethics*, *4*(2), 229-245.

Creswell, J. W., & Poth, C. N. (2016). *Qualitative inquiry and research design: Choosing among five approaches*. Sage publications.

Defensie, M. v. (2020). *Defensievisie 2035*.

Defensie, M. v. (2023). *Defensie Strategie Data Science en AI 2023-2027*.

Defensie, M. v. (2025). *Nederlandse Defensie Doctrine*.

Devitt, S. K. (2024). Meaningful human command: Advance control directives as a method to enable moral and legal responsibility for autonomous weapons systems. In *Research Handbook on Meaningful Human Control of Artificial Intelligence Systems* (pp. 53-80). Edward Elgar Publishing.
Dresp-Langley, B. (2023). The weaponization of artificial intelligence: What the public needs to be aware of. *Frontiers in artificial intelligence*, *6*, 1154184.
Ebrahimi, V. (2024). The Role of Law in Governing Artificial Intelligence in the Context of Global Warfare. *Legal Studies in Digital Age*, *3*(3), 23-30.
Ecemis Yilmaz, H. K. (2023). Autonomous Weapon Systems and the Just War Theory: Challenges and Implications. *Inonu UL Rev.*, *14*, 401.
Ekelhof, M. (2019). Moving beyond semantics on autonomous weapons: Meaningful human control in operation. *Global Policy*, *10*(3), 343-348.
Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and Machines*, *14*, 349-379.
Friese, S. (2019). Qualitative data analysis with ATLAS. ti.
Hammond, D. N. (2014). Autonomous weapons and the problem of state accountability. *Chi. J. Int'l L.*, *15*, 652.
Hughes, J. G. (2020). The law of armed conflict issues created by programming automatic target recognition systems using deep learning methods. *Yearbook of International Humanitarian Law, Volume 21 (2018)*, 99-135.
ICRC. (2011). *Summary of the Geneva Conventions of 1949 and Their Additional Protocol*.
ICRC. (2019). *Autonomy, artificial intelligence and robotics: Technical aspects of human control*
King, A. (2024). Robot wars: Autonomous drone swarms and the battlefield of the future. *Journal of Strategic Studies*, *47*(2), 185-213.
Kreps, S. (2021). Democratizing harm: Artificial intelligence in the hands of nonstate actors. *Foreign Policy*.
Kunertova, D. (2024). *Learning from the Ukrainian Battlefield: Tomorrow's Drone Warfare, Today's Innovation Challenge*.
Kunertova, D., & Herzog, S. (2024). Emerging and Disruptive Technologies Transform, but Do Not Lift, the Fog of War–Evidence from Russia's War on Ukraine. *Russia's War Against Ukraine–Complexity of Contemporary Clausewitzian War*, *2024*, 146-161.
Kwik, J. (2022a). A Practicable Operationalisation of Meaningful Human Control. *Laws*, *11*(3). https://doi.org/10.3390/laws11030043
Kwik, J. (2022b). A practicable operationalisation of meaningful human control. *Laws*, *11*(3), 43.
Kwik, J. (2024). *Lawfully Using Autonomous Weapon Technologies*. Springer.
Lincoln, Y., & Guba, E. G. (1985). *Naturalistic inquiry*. Sage Publications.
Longpre, S., Storm, M., & Shah, R. (2022). Lethal autonomous weapons systems & artificial intelligence: Trends, challenges, and policies. *Edited by Kevin McDermott. MIT Science Policy Review*, *3*, 47-56.
Malterud, K., Siersma, V. D., & Guassora, A. D. (2016). Sample size in qualitative interview studies: guided by information power. *Qualitative health research*, *26*(13), 1753-1760.
Margulies, P. (2017). Making autonomous weapons accountable: command responsibility for computer-guided lethal force in armed conflicts. In *Research handbook on remote warfare* (pp. 405-442). Edward Elgar Publishing.
Molloy, D. O. (2024). Drones in Modern Warfare: Lessons Learnt from the War in Ukraine. *Australian Army Research Centre*.
NATO. (2024). *NATO's revised Artificial Intelligence (AI) strategy*.
Nowell, L. S., Norris, J. M., White, D. E., & Moules, N. J. (2017). Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods*, *16*(1), 1609406917733847.
Oimann, A. K., & Tollon, F. (2025). Responsibility gaps and technology: Old wine in new bottles? *Journal of Applied Philosophy*, *42*(1), 337-356.

Poitras, R. (2018). Article 36 weapons reviews & autonomous weapons systems: Supporting an international review standard. *Am. U. Int'l L. Rev.*, *34*, 465.
Rickli, J.-M., & Mantellassi, F. (2024). The War in Ukraine: Reality Check for Emerging Technologies and the Future of Warfare. *Geneva Centre for Security Policy, Geneva, Switzerland*.
Riesen, E. (2022). The moral case for the development and use of autonomous weapon systems. *Journal of Military Ethics*, *21*(2), 132-150.
Santoni de Sio, F., & Van den Hoven, J. (2018). Meaningful human control over autonomous systems: A philosophical account. *Frontiers in Robotics and AI*, *5*, 15.
Saxon, D. (2024). Military AI and Accountability of Individuals and States for War Crimes in the Ukraine. In *Responsible Use of AI in Military Systems* (pp. 169-191). Chapman and Hall/CRC.
Schwarz, E. (2021). Autonomous weapons systems, artificial intelligence, and the problem of meaningful human control. *Philosophical Journal of Conflict and Violence*.
Seixas-Nunes, A. (2020). Autonomous weapons systems and the procedural accountability gap. *Brook. J. Int'l L.*, *46*, 421.
Sharkey, N. (2016). Saying ‘no!'to lethal autonomous targeting. In *Military ethics and emerging technologies* (pp. 132-146). Routledge.
Smith, P. T. (2022). Resolving responsibility gaps for lethal autonomous weapon systems. *Frontiers in big data*, *5*, 1038507.
Soeters, J., Shields, P. M., & Rietjens, S. J. (2014). *Routledge handbook of research methods in military studies*. Routledge London.
Taddeo, M., & Blanchard, A. (2022). A Comparative Analysis of the Definitions of Autonomous Weapons Systems. *Sci Eng Ethics*, *28*(5), 37. https://doi.org/10.1007/s11948-022-00392-3
Taddeo, M., McNeish, D., Blanchard, A., & Edgar, E. (2022). Ethical principles for artificial intelligence in national defence. In *The 2021 Yearbook of the Digital Ethics Lab* (pp. 261-283). Springer.
Umbrello, S. (2021). Coupling levels of abstraction in understanding meaningful human control of autonomous weapons: A two-tiered approach. *Ethics and Information Technology*, *23*(3), 455-464.
Verbruggen, M., & Boulanin, V. (2017a). Mapping the development of autonomy in weapon systems.
Verbruggen, M., & Boulanin, V. (2017b). SIPRI compendium on article 36 reviews.
Verdiesen, I., Aler Tubella, A., & Dignum, V. (2021). Integrating comprehensive human oversight in drone deployment: a conceptual framework applied to the case of military surveillance drones. *Information*, *12*(9), 385.
Verdiesen, I., Santoni de Sio, F., & Dignum, V. (2021). Accountability and control over autonomous weapon systems: a framework for comprehensive human oversight. *Minds and Machines*, *31*(1), 137-163.
Wagner, M. (2014). The dehumanization of international humanitarian law: legal, ethical, and political implications of autonomous weapon systems. *Vand. J. Transnat'l L.*, *47*, 1371.
Walzer, M. (2015). *Just and unjust wars: A moral argument with historical illustrations*. Basic books.
Wood, N. G. (2023). Autonomous weapon systems and responsibility gaps: a taxonomy. *Ethics and Information Technology*, *25*(1), 16.
Yin, R. K. (2009). *Case Study research: Design and methods* (Vol. 5). Sage

7 Appendices

7.1 Appendix A: List of Interview Questions, in Dutch

Wat is je naam?

Kunt u iets vertellen over uw functie en taken binnen de Landmacht?

Hoelang heb je AWS in je dossier en in welke hoedanigheid werk je samen hiermee?

Kan je mij meer vertellen over hoe het ontwikkel en inzet proces van AWS bij de Landmacht (Defensie)?

- Wat is jouw rol binnen dit proces?
- Is dit een centraal proces?
- Wie/wat is de duwende kracht?
- Wat zijn de belangrijkste stakeholders voor de ontwikkeling en inzet van AWS binnen de Landmacht?
- Hoe werk je samen met deze stakeholders?
- Hoe werk je samen met stakeholders, zoals commerciële bedrijven, buiten Defensie?
    - Stel je hierbij eisen, bijvoorbeeld in hoe algoritmes worden afgesteld?
    - Wordt de training data gedeeld met defensie?

Kan je mij meer vertellen over de data die gebruikt wordt bij de ontwikkeling van AWS?

- Wordt deze data zelf geproduceerd?
- Hoe ‘leert” het systeem?
- Is er inzage in hoe de AWS op bepaalde beslissingen komt?
- Hoe kan je ervoor zorgen dat de training data overeenkomt met de actuele situatie?

Wat betekent voor jou autonomie?

Hoe wil de Landmacht autonome wapensystemen gaan implementeren binnen de operationele troepen?

- In welke context zou een commandant kiezen voor een AWS?
- Wat voor functies kunnen AWS gebruiken?
- Hoe zouden AWS kunnen worden ingezet in de targeting cycle?
    - Kan je mij grofweg uitleggen hoe de targeting cycle werkt, en waar AWS zullen vallen?
    - Deliberate en dynamic targeting?
- Zijn er aanpassingen in bijvoorbeeld Rules of Engagement of Doctrine nodig om dit te bewerkstellen?
- Wat is ervoor nodig om deze implementatie ethisch te laten verlopen?
    - Zijn verantwoordings mechanismes, zoals After Action Report, in te zetten bij AWS?
- Wat heeft een commandant nodig om te beslissen of hij/zij AWS wil inzetten?
    - Hoe kan een commandant ervoor zorgen dat gebruik van AWS proportioneel en noodzakelijk is?
    - Welke context?

    - Betrouwbaarheid van een systeem?

Hoe wil de Landmacht autonome wapensystemen gaan inzetten?

Wat mag een AWS wel en niet?

Welke ethische normen zijn belangrijk bij de ontwikkeling en inzet van autonome wapensystemen?

Hoe worden deze concrete normen omgezet in technische requirements?

- Wat betekent de context voor dit verhaal?

Hoe kunnen mensen de controle houden over AWS?

Wat betekent Meaningful Human Control voor jou?

Hoe ziet human-on-the-loop controle eruit?

- Zijn er mechanismen om hier verantwoording voor te krijgen?

Wordt er verantwoording afgelegd voor (foutief) gedrag van AWS?

- Bij welke instantie wordt er verantwoording afgelegd en hoe gaat dit te werk?
- Wie maakt er een verslag van het (foutief) gedrag van een AWS?
- Kan het duidelijk worden vanuit de logs hoe de AWS zich heeft gedragen?

Hoe wordt er geleerd van inzet van Autonome wapensystemen?

- Is er een feedback loop gebouwd omtrent inzet van AWS?
- Hoe wordt dit gedaan?

Wat zou volgens u nodig zijn om ethisch toezicht op autonome systemen binnen Defensie te verbeteren?

Welke rol ziet u voor opleidingen en trainingen in het waarborgen van ethiek?

Heeft u nog aanbevelingen voor dit onderzoek of zaken die u graag terugziet in het eindrapport?