

ARTICLE

 OPEN ACCESS



Operating itself safely: merging the concepts of ‘safe to operate’ and ‘operate safely’ for lethal autonomous weapons systems containing artificial intelligence

Peter Spayne^a, Laura Lacey^a, Marie Cahillane^b and Alistair Saddington^a

^aCentre for Defence Engineering, Cranfield University, Shrivenham, UK; ^bCentre for Electronic Warfare, Information and Cyber, Cranfield Defence and Security, Cranfield University, Shrivenham, UK

ABSTRACT

The Ministry of Defence, specifically the Royal Navy, uses the ‘Duty Holder Structure’ to manage how it complies with deviations to maritime laws and health and safety regulations where military necessity requires it. The output statements ensuring compliance are ‘safe to operate’ certification for all platforms and equipment, and the ‘operate safely’ declaration for people who are suitably trained within the organisation. Together this forms *the Safety Case*. Consider a handgun; the weapon has calibration, design and maintenance certification to prove it is *safe to operate*, and the soldier is trained to be qualified competent to make predictable judgement calls on how and when to pull the trigger (*operate safely*). Picture those statements as separate circles drawn on a Venn diagram. As levels of autonomy and complexity are dialled up the two circles converge. Should autonomy increase to the point that the decision to fire be under the control of an Artificial Intelligence within the weapon’s software then the two circles will overlap. This paper details research conclusions within the overlap, and proposes a new methodology able to certify that an AI based autonomous weapons system is “safe to operate itself safely” when in an autonomous state.

ARTICLE HISTORY

Received 29 March 2024
Accepted 8 October 2024

KEYWORDS

AI; LAWS; Safety; MOD

1. Introduction

AI based Autonomous systems are gaining significance in various industries, including the United Kingdom (UK) Armed Forces (Roberts 2019), yet their assurance remains inadequately understood. This position paper examines the gaps in assurance for Artificial Intelligence (AI), autonomous systems, and more specifically, lethal autonomous weapon systems (LAWS),¹ given the current regulatory landscape and emerging regulations. Assurance methods used in traditional systems within the Royal Navy and wider civilian industries that use autonomous systems are evaluated. A novel, “Operating Itself Safely” theoretical assurance framework is proposed for the military application of AI based LAWS. The framework addresses the unique challenges presented by

CONTACT Peter Spayne  peter.spayne@cranfield.ac.uk  Centre for Defence Engineering, Cranfield University, Shrivenham SN6 8LA, UK

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

autonomous systems and more specifically LAWS. It explores the convergence of the traditionally distinct concepts of “safe to operate” and “operate safely.”

Advancements in technology have afforded command elements enhanced visibility and command over the entire modern battlespace. However, discussions within the Group of Government Experts (GGE) have emphasised that military organisations do not endorse full autonomy of weapons. This perspective aligns with the UK Ministry of Defence’s (MOD) stance (Ministry of Defence 2021), which abstains from pursuing the development of LAWS. However, due to the nature of modern warfare and the increasing use of semi-automated systems in live operations, detailed analysis of the technology is still required.

The literature review that follows considers existing assurance methods; analysing those for traditional machinery, military applications, and autonomous systems in non-military contexts. A recurring theme is the convergence of the concepts “safe to operate” and “operate safely” in these methods. This review identifies common limitations and sets the stage for the proposed merger of these two concepts.

2. Current regulations and requirements

Formal regulations for LAWS are lacking due to the immaturity of the technology (Bills 2014; Human Rights Watch 2023). However, this landscape is changing at pace due to the recent outbreaks of global conflict. Legal working groups have formed with the aim to provide user control for new technology, ensuring reasonable certainty and proportionality of the use of LAWS in line with International Humanitarian Law (IHL) (Boulainin and Verbruggen 2015; Boulainin, Davidson, Goussac, and Carlsson 2020).

The NIST (2023) study leading to the AI Risk Management Framework (RMF) highlights specific AI risks to be considered alongside traditional risks. These new risks are present in the case of LAWS technology that is based on AI and can impact a broader audience and have far-reaching consequences. The RMF checklist covers the following additional risks and opportunities that must be addressed in their bespoke AI-Safety Case:

- Data may not accurately represent the intended AI system environment. AI systems rely on complex training data, necessitating risk assessment. Changes to training data or requirements during training impact AI system performance.
- Datasets for AI training can lose context or become outdated.
- Assessing all risks in large, complex AI systems is impossible.
- Pre-trained models enhance performance but introduce uncertainty and bias.
- Predicting failure modes in large-scale, pre-trained models is challenging.
- Privacy and security risks from data aggregation require individual assessment.
- AI systems require frequent maintenance, often triggered by data or model changes.
- Identical systems may not reproduce identical software models due to learning.
- Traditional software testing standards are insufficient for AI system assessment.
- Regular AI software testing is difficult due to different controls.
- Computational costs and environmental impact of AI model development are unknown.
- Detecting side effects of AI systems is challenging for risk management.

To begin developing regulations for LAWS the Stockholm International Peace Research Institute (SIPRI) conducted research identifying technical, legal, operational, and ethical issues with autonomous weapons. It focused on human control parameters and offered regulatory options for emerging technology (Boulainin, Davidson, Goussac, & Carlsson 2020).

Since the first SIPRI papers were released, the European Commission (EC) has begun planning a regulatory framework for AI systems, with LAWS potentially included as they contain AI (European Commission 2022). The framework is based on the following four risk levels to address unpredictable outcomes, aiming to protect individuals from unfair prejudice by ensuring transparency in AI decision-making:

Unacceptable risk systems (e.g., social scoring systems as depicted fictionally in Netflix *Black Mirror*: Nosedive episode (Booker 2016) would be banned due to threats to life or human rights.

High-risk systems, such as AI controlling critical infrastructure, law enforcement, or employment, would operate under strict rules and generally be prohibited, except under exceptional court-authorised circumstances.

Limited-risk systems like chatbots must have clear user-awareness mechanisms for informed decision-making.

No-risk systems, such as AI in video games, would remain unrestricted.

By encompassing risk mitigation for AI applications, and identifying high-risk applications, stringent requirements for systems and defined obligations for users and providers can be written, with conformity assessments assuring compliance prior to market entry and the establishment of governance structures at both European and national levels (European Commission 2022). It is unclear where LAWS will fall within these categories, as they are designed to take life by their very nature. However, as these regulations are primarily concerned with the safety of users and interacting individuals, they may partially fall under regulations that protect operators, whilst deliberately overlooking the actual purpose of the systems (European Commission 2022).

LAWS must also sit within the IHL framework based on its capability. Challenges and constraints on LAWS within the IHL framework primarily revolve around key principles, including distinction of targets, proportionality, and precautions (Crootoof 2022).

Traditionally, human weapon operators (except for the case of land and sea mines) have been responsible for determining lawful targets and adhering to rules of engagement (Ministry of Defence 2013). These human judgments often involve qualitative and quantitative subconscious assessments. In contrast, AI-based targeting systems operate on mathematical algorithms to identify criteria that satisfy pre-defined rules, as discussed by the Swiss government in their paper “A ‘compliance-based’ approach to autonomous weapon systems” (Swedish Government 2020).

The divergence between human and AI assessments of threat stems from several factors, raising important questions about the applicability of IHLs proportionality rules. Humans rely on a combination of contextual understanding, ethical judgment, and emotional intelligence when assessing threats, which allows for nuanced decision-making that considers the broader humanitarian impact. In contrast, AI systems base their assessments on pre-programmed algorithms and training data, which may lack the

ability to interpret complex human behaviours and the subtleties of a dynamic battlefield. AI systems, acting autonomously, may make decisions without considering the potential for wider civilian harm, loss of life, injury, or damage to civilian objects in advance. Similarly, assessing potential military advantage (on operations where lethal force is likely) in numerical terms becomes challenging, as these variables lack quantification.

IHL rules, designed to prevent excessive civilian harm, traditionally necessitate a deeper level of human judgment that encompasses moral and ethical considerations across various contexts. This reliance on human judgment is a distinctive feature of IHL. AI technology, found within LAWS, is widely used for statistical analysis. At lower levels of autonomy it aids human decision-making by providing insights. However, in such cases, human judgment serves as the final scrutiny before any action is taken. In this context, humans remain in control of ensuring the *proportionality* of decisions. Experts concur that AI processes alone do not fulfil the IHL proportionality assessment, and they cannot replace human decision-making in this regard (ICRC (2014); Thurnher and Schmitt, and Thurnher 2013). Indicating a mapping difficulty between civilian and military uses of AI.

The distinction between a LAWS initial deployment and the subsequent weapons function, which occur under uncertain circumstances (e.g. target, location, timing, and surroundings), makes it challenging for the operator to have complete control over a proportional action. The potential unpredictable nature of the AI's responses and the potential collateral damage further highlight the limitations of relying solely on AI to conduct a proportionality assessment.

Article 49(1) of Protocol 1 of the Geneva Convention defines acts of violence against the adversary whether in offence or defence and is backed up by Articles 51 and 57 relating to the protection of victims of international armed conflicts. These rules provide context for an attack using LAWS (Article 49 - Definition of Attacks and Scope of Application. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I), (1987).

How legal obligations are met when using LAWS will depend on the scope of its use, including the duration of an attack. This means it is critical to define when an attack starts and finishes. Identifying these moments in time frames the legal obligations. In the case of LAWS, "attack" may be read differently to traditional weapons. Boothby (2019) states that an attack begins when LAWS selects a target. This means that even if the weapon (in the case of LAWS) does not function and there are no mortal consequences of being targeted, the attack would still be relevant (Boothby 2019; Sparrow 2015). This opinion correlates with a Red Cross documented commentary on mines from 1982 that states "*there is an attack whenever a person is directly endangered by a mine laid*" (Asaro 2012). Analysis of this correlation focuses on the legal framework surrounding the concept of an "attack" regarding LAWS. Highlighting the necessity of precisely defining the start and end of an attack to establish the corresponding legal obligations. Crucial for applying IHL effectively.

For LAWS to meet IHL requirements, operators must make context-specific, time-bound decisions before deployment to frame a specific operating environment. However, operators need the flexibility to adjust LAWS parameters as circumstances change, and the ability to abort an attack. Boulanin, Davidson, Goussac, and Carlsson (2020)

highlights the importance of users' assumptions about the target's context and environment for the lawfulness of an attack. As changes can occur post-deployment control over the context of deployment is crucial for reliable planning and legal operation.

IHL prohibits indiscriminate weapons, emphasising proportionality and precautions. To limit a LAWS destructive power, users must also predict how LAWS may function in different scenarios. However, this is challenging for a LAWS, which may apply force at unpredictable times and locations. Furthermore, the environment may change after deployment, affecting target legitimacy. Predictability is therefore vital for legality.

In the 2017 Asilomar conference (Conn 2017), the United Nations called for research and policy input to ensure that new autonomous systems consistently reflect human values. They established 23 AI principles, emphasising safety with an aim to avoid arms races whilst upholding human values in AI development. The principles encompass Research, Ethics and Values and Longer-Term Issues. These principles are similar to the AI risks identified by NIST and have been mapped into this assurance proposal.

3. Assurance of traditional technology within the royal navy

This section explores the historical separation of “safe to operate” and “operate safely” in various contexts to establish where the knowledge gap of this research lies. It also highlights compelling reasons for merging these concepts within the realm of autonomous systems. The proposed merger not only harmonises two essential aspects of assurance but also streamlines the approach to system safety.

The Royal Navy has a well-established process for ensuring the safe operation of technology, aligning with the guidance from the UK's Health and Safety Executive (HSE). The procedures for adhering to this guidance in a military context, particularly when operating warships and weapons, are detailed in Defence Safety Authority – Book 2 – Defence Maritime Regulations (DSA02-DMR) (Defence Maritime Regulator 2023). Macro-level guidance is further specified in subordinate publications.

The entire risk management process, encompassing people and machinery, hinges on a complex web of HSE, International Maritime Organisation (IMO), equipment-specific International Standards Organisation (ISO) regulations and others, as illustrated in Figure 1. Given the inherent dangers of military operations, which sometimes involve deliberately causing harm, deviations to HSE guidance to facilitate military operations and actions by service personnel are overseen internally by a system of duty holders (Maritime Regulator 2016).

The Defence Safety Authority (DSA) aims to ensure the safety of equipment and personnel for health, safety, and environmental protection, striving to keep accepted risks “as low as reasonably practicable” (ALARP).

The DSA series of publications enables defence agencies to balance workplace health and safety legislation compliance with the inherently risky nature of military activities. As the military differs from conventional workplaces, and some activities involve causing harm through lethal force. DSA manages safety where it can be applied, such as the training environment. Directive DSA01–1 does not negate the obligation to adhere to UK health, safety, and environmental laws or policies – which the military opts to do (Defence Maritime Regulator 2018).

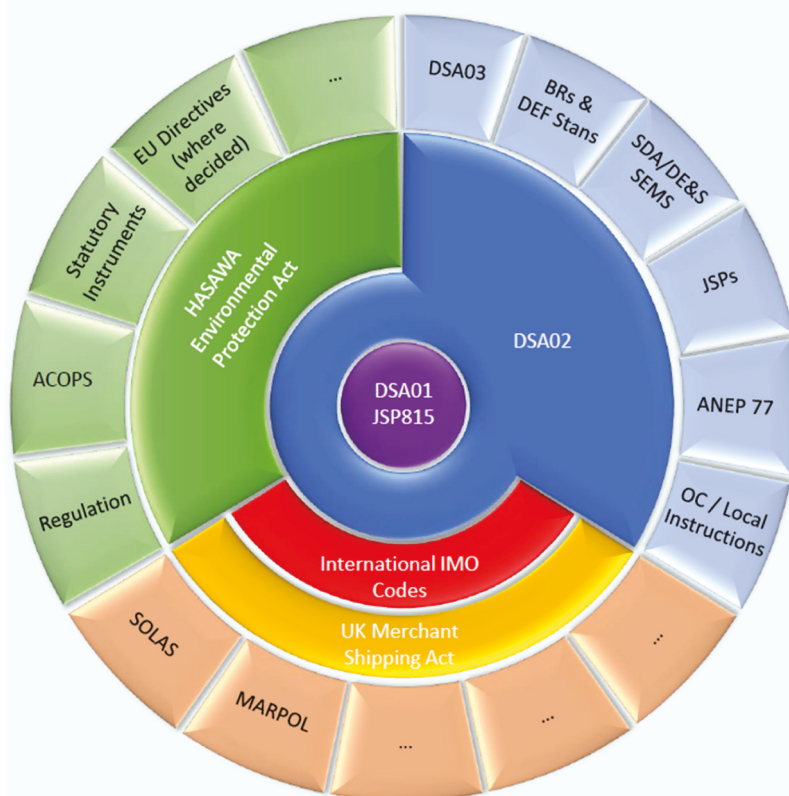


Figure 1. Safety and environmental legislation and regulation hierarchy. Atkins, 2023.

In defence activities, guidance is established to maintain ALARP risk levels in routine operations, with sub-organisations that deliver these functions being deemed “duty holder facing” (Defence Maritime Regulator 2018). Duty holder facing organisations ensure equipment, services, and infrastructure align with agreed safety criteria and promptly report any deviations to the relevant duty holder. (Note, the defence definition of a duty holder differs from the HSE definition (Ministry of Defence 2022)).

To establish *safe to operate*; platform authorities and equipment authorities are tasked with ensuring that equipment, both individually and within a system, is safe. This is accomplished through making compliance arguments in a safety case underpinned by bodies of evidence to achieve certification from the Naval Authority. The Naval Authority establishes the rules of maritime compliance and minimum build standards by reviewing maritime regulations. They combine these regulations into a military-ready product, considering necessary deviations from commercial regulations.

Assurance for equipment operates on a graduated approach, addressing components, systems, and systems of systems. Equipment Authorities manage the introduction of equipment into service, developing safety cases for safe operation and defining integration, training, maintenance, and eventual disposal requirements within larger systems. Once all the necessary evidence is gathered and established, certification is issued to ensure the system’s reliability.

All the equipment on a platform - under the command of a single Commanding Officer (CO), shall be operated in line with governance provided by the Platform and Equipment Authorities.

During high-risk tasks, with potentially significant consequences, a Go/No-Go decision is necessary to proceed. The individual with the duty of care (an obligation to act in a certain way, to certain standards), typically the CO, makes this decision as the management of risks to life (RtL) falls within their authority and legal tolerance. If the required activities generate a greater RtL than they can manage then decision-making is escalated up the chain of duty holders until it reaches an individual with suitable seniority to hold the risk.

Risk taking and deviance from safe practice is evaluated in various situations against the operational benefit. In rare cases, for military advantage, very high risks beyond the acceptable limit may be authorised. For instance, allowing helicopter operations in extreme weather to rescue mariners from a sinking ship.

In exceptional situations that demand immediate actions outside established rules and guidance, operations such as this can proceed under the immediate direction of the CO (who will be held to account should it go wrong), based on a local, informal, risk assessment.

To certify that personnel are *operating safely* when using the equipment, the training authority ensures that all personnel are properly trained and vetted to operate the platform and equipment safely. The CO through their representative committees assure that people behave in accordance with their training.

This current system, as detailed in the DMR series, works well for conventional technology. It includes checks for educational competence and background, access to healthcare, fitness, welfare facilities, and the provision of clothing, protective equipment, and necessary tools. Personnel are compensated for their work, and there are mechanisms in place to acknowledge compliance and non-compliance with the established safety guidelines.

Figure 2 depicts the hierarchy of legislation and documentation starting with the Health and Safety at Work act and percolating through various levels of direction and instruction to ensure that the person on the ground is protected by the correct policy and procedures to keep them safe no matter what their task. The central part of the figure depicts the legislations that duty holders use to conform with legislations as closely as practicable without compromising military operational capability.

Individuals are *operating safely* when they have undergone rigorous training, supervision, and follow provided guidelines. Personnel receive the necessary protective equipment, tools, and documentation and are tested for their competency. A second layer of assurance ensures training and environments are correct, along with assessments of support services like healthcare. An example of this could be the Crew of a warship. Or a single soldier operating a rifle.

However, individuals retain free will, and unexpected behaviours can still occur, as evidenced by high-profile incidents that often appear in the media. Whilst individuals themselves can never be fully assured as safe, evidence from their performance can be used to build the safety case around their suitability to work within the organisation.

Through analysis of the Royal Navy's discipline and promotion regulations it can be concluded that (in the context of safety); when individuals consistently follow what they

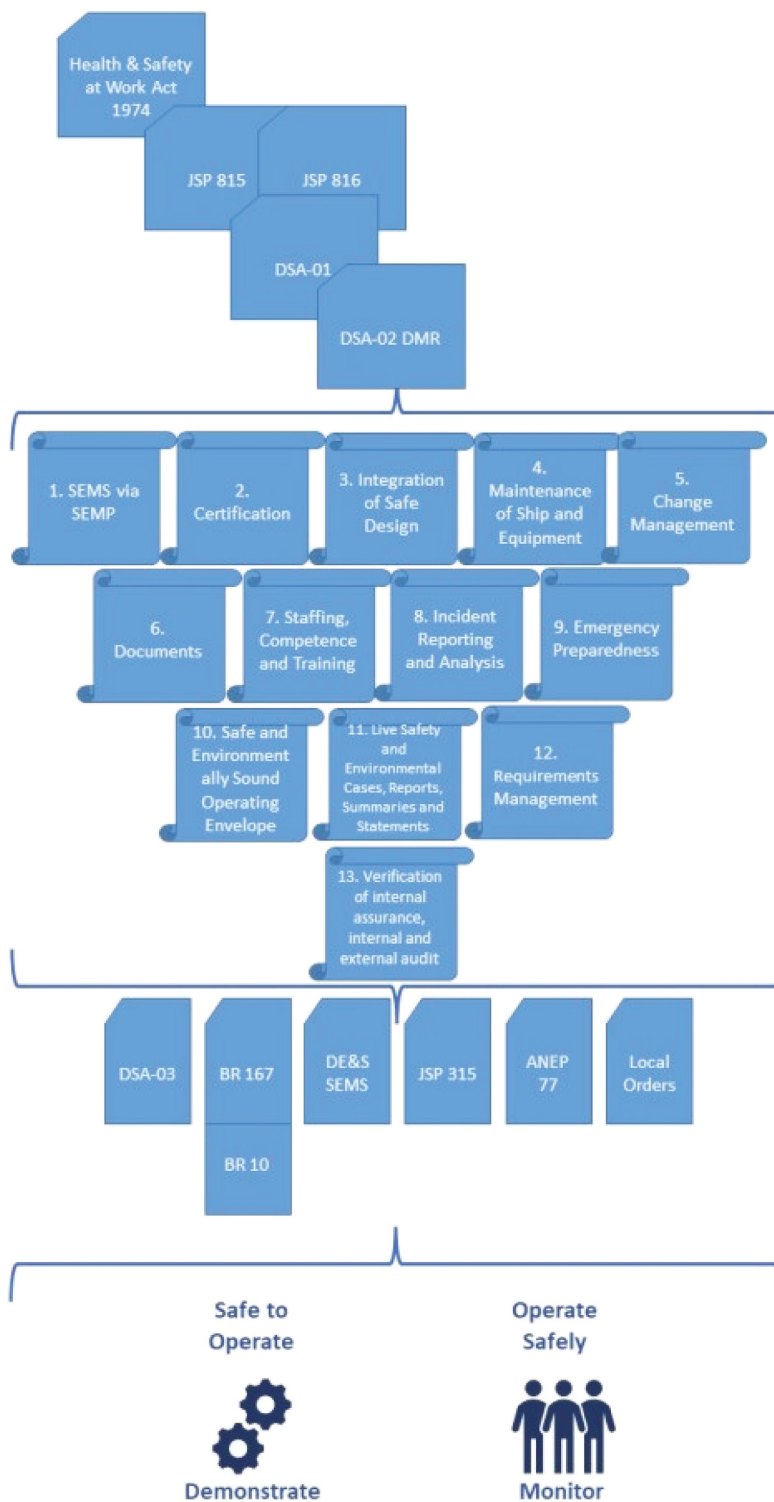


Figure 2. Safety hierarchy. Spayne, 2023.

have been taught and demonstrate trust over time, they are recognised through rank, visually indicated by symbols. Compliance with regulations is rewarded, while defiance results in financial penalties, loss of status, and legal consequences (Royal Navy 2019).

The approaches detailed above regarding machinery are perfectly acceptable in the case of manual, automatic and automated systems, set up or operated by people following documented processes and procedures, as every possible operation and failure can be predicted and mitigated against, assuming correct operation in accordance with the procedures are followed. Should something unexpected occur then the system would be quickly withdrawn for use pending further analysis and repair. This methodology will not work, however, for AI systems that operate themselves, as unknown inputs cannot be mapped or have their risks analysed.

For AI based systems, especially those used in critical applications (LAWS), the unpredictability of inputs and the complexity of their decision-making processes mean that traditional risk mitigation strategies are insufficient. Unlike manual or automated systems, AI operates based on patterns and data it has been trained on, and can encounter novel situations that were not anticipated during the training phase. Whilst similar to a human encountering a situation they were not trained for; this unpredictability necessitates new approaches to safety and reliability due to the lack of human to hold to account post event. The ability to assess AI Risks and then audit and explain AI decisions becomes crucial to ensure accountability and trust in LAWS, particularly when human lives are at stake. Therefore, a third concept, ensuring the ethical use and rigorous testing of AI in autonomous systems is essential to mitigate the risks posed by their autonomous nature.

4. Autonomous assurance in the wider industry

Assurance seeks to establish confidence in autonomous systems, but their capacity to respond in unknown ways to unforeseen situations raises the challenge of pre-emptive assurance. Traditional industry assurance relies on Testing, Evaluation, Verification, and Validation (TEVV) during development, yet autonomous systems, including LAWS, demand continuous runtime assurance. This commitment incurs both temporal and financial costs, driven by the necessity for extensive evidence in support of safety-critical decisions. The assurance level is intrinsically linked to the system's vulnerability to failure and involves ensuring adherence to human and AI ethical values, spanning transparency, fairness, non-maleficence, responsibility, and privacy.

As per Leslie (2019), AI assurance commences with Process-Based assurance, which delves into system development and Product-Based assurance, encompassing TEVV testing. While foundational, autonomous systems require additional principles tailored to their diverse applications.

The creation of software and AI models, whether manual or machine-generated, hinges the responsibility for problem solving onto the programmers. Often, understanding the operational environment that the final system will deploy in is far outside of their remit, prompting collaboration between engineers developing autonomous systems and scientists across varied disciplines to establish governing rules for system behaviour, in consultation with environmental experts, such as soldiers.

Machine learning (ML) when present in LAWS underpins AI by providing the computational techniques and algorithms that enable AI systems to learn from data, adapt, and improve over time without being explicitly programmed plays a pivotal role in facilitating system learning and experiential growth (Kester 2016). All software, irrespective of its origin, must undergo checks to identify errors and unintended outputs. The establishment of requirements guides system design, particularly important since LAWS frequently operate in uncertain settings, resulting in vague and challenging-to-verify requirements. In traditional verification models, testing gauges product compliance against specifications, simplifying assurance. However, the intrinsic ambiguity of autonomous systems allows for individual isolated bodies of evidence to be gathered but complicates the compilation of a comprehensive safety case (Jobin, Ienca, and Vayena 2019).

The reliability of hardware now becomes pivotal, as faulty sensors or uncalibrated devices can introduce erroneous or hazardous outcomes. The speed of system operation is critical, as even minor lapses in latency, (milliseconds), could yield fatal consequences (Leslie 2019).

Due to complex third-party commercial agreements and a vast web of intellectual property rights surrounding different software licences that may all make up a single system, the accessibility of design documentation and specifications for non-commercial equipment can pose challenges. This introduces a risk that users may “go rogue” developing undocumented front-line solutions and crafting custom test routines. While these routines instil confidence in isolated system aspects on the ground, they lack actual verified environmental data. In cases where live data is used, continuous assurance should incorporate software-based self-testing to affirm the system’s alignment with its design standards (DSTL 2021).

Regarding training data, its suitability for instructing machine learning models to correlate inputs to suitable outputs must be assessed. Specific requirements, such as being able to identify and designate a “friendly tank” or “speed bump,” are critical in instructing the system on input data processing and general behaviour learning. A key concern is preventing unintentional biases in data, which can arise from various sources such as sampling, labelling, or measurement biases. These biases can occur due to narrow sample groups, faulty sensors, or feedback loops where biased AI outputs perpetuate and amplify the biases present in the original training data. As the system’s behaviour mirrors the training data, trivial errors can be carried through and develop into significant problems during later deployments. To assure against this, data validation is required, utilising independent training sets or synthetic data resembling the expected operational environment – although synthetic data can sometimes produce difficult to emulate edge cases. Care should be taken as poorly formatted test data, which may have been gamed to ensure the systems success, can lead to false system confidence. This often necessitates third-party oversight to prevent this from occurring (DSTL 2021).

The assurance of autonomous systems algorithms in isolation depends on the various algorithm types, as each dictate a different nature of assurance. Supervised learning processes labelled data for predicting outcomes, whilst unsupervised learning processes unlabelled data to extract patterns. Semi-supervised learning combines both labelled and unlabelled data for processing, and reinforcement learning improves through trial and error, modifying actions based on outcomes. Algorithm reliability hinges on defining its

role, and prioritises the prevention of specific failures over exhaustive output specification. The communication of system-level requirements to the algorithm's function does not map well to traditional hierarchical drawings and also requires verification by an independent team, through extensive testing (Leslie 2019). Ultimately, the assurance process evaluates whether the algorithm aligns with system requirements.

Live inputs are derived from operational use, subsystem failures, or deliberate attacks, and vary in their extent of coverage. Algorithm testing typically transpires within synthetic environments. The careful calibration of testing to avoid overtraining, which can hinder real-world performance, is central to a balanced assurance case. The algorithm serves as the bridge from input to output within the system, interfacing with a fusion of sensors and subsystems. Integration features encompass backup sensor use, input analysis and safety mode activation to ensure system safety during subsystem failures (Leedeo Academy 2020).

Unexpected failures can occur when the algorithm receives data beyond its training, resulting in unpredictable outputs, much like a human when facing something unexpected. Analysing the outputs generated because of unexpected inputs in a synthetic or safe live environment is key to detecting such failures. Safety software may act as a preventative measure to prevent dangerous predefined outputs. Simultaneously operating multiple identical algorithms and comparing their outputs to identify rogue code is a common practice in flight software (DSTL 2021) and an accepted method to identify defective ML models. Self-monitoring functions provide an additional safeguard by substituting incorrect input values (when identified) with safe defaults. While monitoring can reduce the need for extensive pre-deployment assurance, balance is essential to prevent designing an overly cautious system that could become permanently immobile. Self-monitor features also record system activities for debugging, future development, and accident investigation (DSTL 2021).

Comprehending diverse environments, both at domain level and technical, represents a formidable challenge. Identifying the intended operational environment (and remaining within it) is a critical component of the assurance case (that can quite literally change with the weather) and encompasses the self-identification of a system operation outside its intended parameters and then recording output responses in such instances. Human behaviour, be they users, pedestrians, or adversaries, significantly influence the operating environment. Accommodating pedestrians with lower risk tolerance and facilitating intuitive user handover points for the assumption of full system control are recognised as essential considerations (Leslie 2019).

A definition of *safe* is paramount to identifying reactions to unexpected scenarios that are unwanted or dangerous. Once environmental and safety definitions are established, the design and assurance of safe behaviours become central. While binary behaviours like “the system must” or “the system shall never” are important, many contextual behaviours necessitate a case-by-case approach (DSTL 2021).

For example, in the case of a vehicle conducting a search and rescue mission, environmental complexity such as unpredictable weather conditions, variable terrain, and the presence of obstacles like debris or unstable structures can challenge the system's safety and operational decisions. The AI must adapt its behaviour to ensure the safety of both the rescuers and the victims, avoiding areas where its actions could cause further harm or destabilise the environment.

In a more military application, the complexity of the mission scenario increases with factors such as enemy tactics, civilian presence, and dynamic battlefields. An AI operating a drone, for instance, must distinguish between combatants and non-combatants in urban warfare, where buildings can obscure vision and civilians can be interspersed with enemy forces. The drone's decision-making needs to be contextually aware to minimise collateral damage and comply with IHL.

Wider environmental complexities such as electromagnetic interference or cyber-attacks also affect the reliability of sensors and communication systems. In such cases, the AI needs to recognise anomalies in data and adjust its operations to maintain safety. For example, a self-driving fighting vehicle encountering sensor blindness due to heavy rain or fog must modify its driving behaviour to prevent accidents, potentially by slowing down or pulling over until conditions improve.

Behaviours outside of the deployed context encompass activities such as maintenance, training, storage, or transport also demand attention to ensure secure shutdown and protection against adversarial attacks. System-level assurance hinges on the understanding of the desired system behaviours within specific operating contexts and whether the system aligns with the operator's intentions (DSTL 2021).

In practice, systems seldom operate in isolation; they often interact with humans and other machines, introducing significant complexity to their assurance frameworks. LAWS can exhibit rapid, predictable responses within tightly confined operating envelopes but often lack broader environmental awareness, impacting their decision-making processes and increasing the risk of unintended consequences. Interactions between human-machine and machine-machine teams introduce unknown variables, complicating assurance and reliability. While some systems are designed to work in concert, others coexist by circumstance, further expanding the scope of potential inputs and interactions. Addressing the unexpected outcomes of mutual interference at the point of design, especially when systems lack designed compatibility, becomes a critical consideration (Leslie 2019). For instance, in urban warfare, LAWS might struggle to differentiate between combatants and non-combatants due to limited situational awareness. Effective risk mitigation strategies, such as enhanced interoperability standards and robust testing procedures, are essential to ensure safe and reliable operation in these complex environments.

Considering LAWS' operation in potentially adversarial settings, protection is imperative. Non-combat vulnerabilities, such as "back doors" created by malicious data or changes in the operating environment, pose significant threats. For instance, during a conflict, adversaries might exploit these vulnerabilities to disrupt or manipulate LAWS operations. The principal safeguard against such threats is the maintenance of high-security standards for training data, ensuring it originates from reliable and trustworthy sources and adheres to the principles of trustworthy AI (Li et al. 2023). Additionally, employing multi-layered security approaches, real-time monitoring, and robust incident response plans are crucial for comprehensive protection. For example, the implementation of secure data pipelines and continuous security audits can help mitigate risks and enhance the resilience of LAWS in complex environments. These measures, combined with rigorous adherence to international security standards, ensure that LAWS operate safely and effectively, even under adversarial conditions.

4.1. Conducting assurance

Autonomous systems are structured across three levels, as depicted in [Figure 3](#). However, this layered approach can be perplexing, as each layer may appear to function as an independent autonomous system (Safety of Autonomous Systems Working Group [2022](#)).

Layer 3 (System Level): The outermost level, often termed the “System Level” or “Platform,” represents the final autonomous system as a product, be it a physical entity or software. This level is defined by its ability to interact with its environment directly.

Layer 2 (Autonomy Architecture): This level addresses the autonomy software within subsystems, with the potential for multiple subsystems within a system. It concerns components, not the entire system.

Layer 1 (Computation): Encompasses individual processes that translate inputs into outputs. Complex systems typically involve numerous architectures and computations at this foundational level.

The Safety of Autonomous Systems Working Group ([2022](#)) presents a summary of data collected from various sources, such as the ‘Assuring Autonomy International Program (Aaip), Assuring Machine Learning for use in Autonomous Systems (AMLAS), UL4600 Autonomous Vehicle Safety Standards, The Organisation of Economic and Cooperative Developments (OECD) Principles for AI and the Assessment lists for Trustworthy AI (ALTA) frameworks. Their research served as a reference point for the creation of safety cases for autonomous systems, running parallel to traditional safety cases for regular technology. Helping to understanding the depth and interdependencies within autonomous systems.

The NIST ([2023](#)) study leading to the AI Risk Management Framework (RMF) highlights specific AI risks to be considered alongside traditional risks. These risks can impact a broader audience and have far-reaching consequences. The RMF checklist covers the following additional risks and opportunities that must be addressed in their bespoke AI-Safety Case:

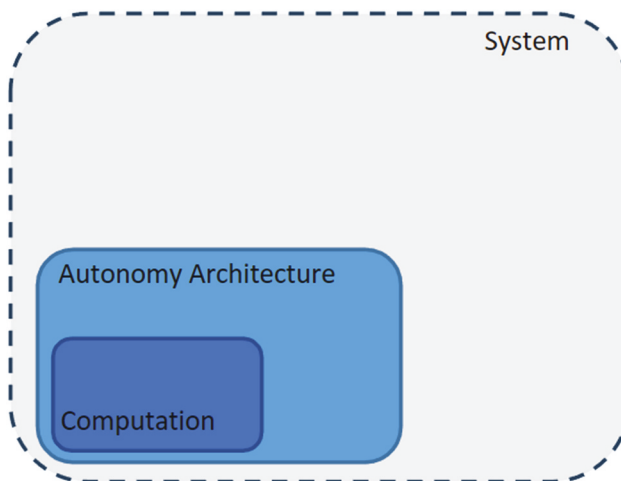


Figure 3. Autonomy framework. Spayne, 2023.

- Data may not accurately represent the intended AI system environment. AI systems rely on complex training data, necessitating risk assessment. Changes to training data or requirements during training impact AI system performance.
- Datasets for AI training can lose context or become outdated.
- Assessing all risks in large, complex AI systems is impossible.
- Pre-trained models enhance performance but introduce uncertainty and bias.
- Predicting failure modes in large-scale, pre-trained models is challenging.
- Privacy and security risks from data aggregation require individual assessment.
- AI systems require frequent maintenance, often triggered by data or model changes.
- Identical systems may not reproduce identical software models due to learning.
- Traditional software testing standards are insufficient for AI system assessment.
- Regular AI software testing is difficult due to different controls.
- Computational costs and environmental impact of AI model development are unknown.
- Detecting side effects of AI systems is challenging for risk management.

The Artificial Intelligence High Level Expert Group (AI HLEG) in 2020 created the ALTA framework. This guidance allows developers to self-assess their products during development, focusing on technical and social aspects. ALTA covers seven areas, including human agency, oversight, technical robustness, privacy, transparency, diversity, societal and environmental wellbeing, and accountability (High Level Expert Group on Artificial Intelligence 2022).

The human agency aspect considers AI's influence on human behaviour and perceptions. Oversight involves classifying the system, determining human interactions, and recommending training, detection methods, and monitoring. Technical robustness focuses on the system's ability to adapt to unexpected changes and withstand malicious interactions. It includes resilience to attacks, cybersecurity, and measures to safeguard against cyber threats. Privacy assessment is vital for respecting human rights and ensuring mental and physical integrity during system interactions. Transparency involves traceability, explainability, and open communication about system limitations. Stakeholder participation is essential, with feedback sought throughout the system's lifecycle. Social wellbeing involves considering the effects on society and individuals due to AI interactions.

4.2. Comparison of wider industry standards

After a detailed comparison of regulations, legislations and standards across a multitude of industries that included ISO 26262 ASIL, IEC 61508 SIL, ARP 4761 DAL ISO 13849, IEC 62304 and various chain methods, it was noted that the ASIL method from ISO 26262 used primarily for the automotive sector had the best promise for adaptation to autonomous systems due to its three-dimensional risk assessment that could be tailored for differing maturities of AI. This would work well merged with IEC 61508 SIL for general safety-critical systems taking influence from ARP 4761 methodologies for aviation. While these standards traditionally serve different industries, they share common principles in safety assurance and would work well merged for autonomous applications within the Royal Navy and MOD.

Various safety assessment approaches exist and there is no definitive *best practice* as each is tailored towards a specific industry, the comparison conducted in this investigation was not exhaustive. Observations concluded that ASIL and SIL assess a hazards' impact, while ARP 4761 aligns with SIL principles in aviation. Markov Chains offer insights into complex system failures and make for a very useful post-accident investigation tool. ISO 13849 evaluates system performance and mitigation strategies (International Standards Organisation 2015), while IEC 62304 prioritises "safe by design" approaches to minimise safety assessments (Hall 2010). Each industry champions its safety standard as the definitive one despite them all operating differently. Ultimate choice for use with autonomous systems depends on the system component type and safety philosophy.

In contrast, the Royal Navy's current two-dimensional risk matrix as detailed in BR10 evaluates likelihood and severity for human and machine safety independently. Of greatest value is ASIL in ISO 26262 that adopts a three-dimensional approach and introduces the capability of a controlling agent. This recognises autonomous systems' unique challenges, emphasising their need to operate safely without direct human intervention. The 3D risk assessment model considers safe operation, self-assessment, and real-time decision-making for safety.

ASIL specifies functional safety management, design, verification, validation, and relationships between manufacturers and operators (Hobbs, and Lee 2013).

Traditional risk management matrices used by the Royal Navy multiply likelihood by consequence. ISO 26262 ASIL is different, however, with Severity (S), Exposure (E), and Controllability (C) as detailed in Table 1. The equations are:

$$\text{Traditional Risk: RISK} = \text{SEVERITY} \times \text{LIKELIHOOD}$$

$$\text{ISO 26262 ASIL: ASIL} = \text{SEVERITY} (\text{EXPOSURE} \times \text{CONTROLABILITY})$$

The ASIL matrix (Table 1) (International Standards Organisation 2011) shows the risk of potential outcomes in case of malfunctions in automobiles. Whilst originally intended for traditional technology risk analysis and safety system design, it aids stakeholders in understanding the recoverability of malfunctioning autonomous systems, such as those impacting safety cases for current vehicle subsystems. Autonomous subsystems common in automobiles include lane-holding technology or adaptive cruise control promote ASIL-D compliance (Mobility Insider 2020). Table 2 showing the current MOD risk assessment matrix is detailed for comparison.

ISO 26262 classifies ASIL-D as the highest hazard level, requiring the most rigorous safety measures. While, Quality Management (QM), one level below ASIL-A, implies that all risks are tolerable, exempting the need for safety management at all.

For example, in a scenario where a vehicle suffered a catastrophic tyre failure (blow-out) whilst traveling at high speed, the traditional risk matrix, like the one used by the Royal Navy, would assess risk based on likelihood and severity. In doing this many over cautious assumptions would need to be made. In contrast, the ASIL approach, (Table 1), considers Severity as usual, but it also factors in the operating environment, running hours, local population density and more (exposure). It weighs these factors against the

Table 1. ASIL MATRIX (ISO 26,262).

Severity	Exposure	Controllability		
		C1 Simple	C2 Normal	C3 Difficult
S1 Light Injury	E1 Very Low	QM	QM	QM
	E2 Low	QM	QM	QM
	E3 Med	QM	QM	A
	E4 High	QM	A	B
S2 Life- Threatening Injury	E1 Very Low	QM	QM	QM
	E2 Low	QM	QM	A
	E3 Med	QM	A	B
	E4 High	A	B	C
S3 Fatal	E1 Very Low	QM	QM	A
	E2 Low	QM	A	B
	E3 Med	A	B	C
	E4 High	B	C	D

Table 2. Typical risk management matrix (MOD form 5010).

High (3)	Med 3	High 6	High 9
Medium (2)	Low 2	Med 4	High 6
Low (1)	Low 1	Low 2	Med 3
Likelihood x Severity	Low (1) Minor Injury	Medium (2) Serious Injury	High (3) Fatality

*Multiple risk assessment matrices are in use in the MOD, that add greater complexity.

ability of an average driver to regain control (controllability) and ensure safety. Thus, the safety case can consider the type of road, skill and awareness of the driver.

This ASIL method can lower the risk level for a potentially life-threatening event if it is deemed to be highly controllable by a skilled operator (C1) This, therefore, avoids the need for excessive and costly safety measures. When reapplied as an *Autonomy* Safety Integrity Level, the approach factors in controllability, the assured environment and exposure to people nearby. It helps predict the severity of an unexpected outcome and defines necessary controls to mitigate it.

For LAWS assurance, consider a sensor failure. If there are adequate warnings and software able to facilitate a safe recovery, the system can ensure safety as long as the recovery task remains at a C0 or C1 level, which can be handled by a skilled operator or AI, either manually or remotely. The ASIL system can be tailored to allow the system to initiate higher-level safety measures when no operator is present.

4.2.1. *Integration of wider considerations*

To ensure the safety of autonomous systems, further incorporating AI-specific considerations of a safety case into the traditional TEPIDOIL framework (Ministry of Defence 2009) (expanded below) is vital. The Royal Navy, in particular, employs the TEPIDOIL structure for procurement, which takes into account the organisation's responsibility to invest in wider aspects of the system's additional infrastructure requirements of 'Training, Equipment, Personnel, Information, Doctrine, Organisation, Infrastructure and Logistics. However, the integration of AI into technology brings new risks and challenges that require further analysis.

By expanding the TEPIDOIL framework to include wider AI-related factors, organisations can address the unique risks associated with AI in autonomous systems. This holistic approach enhances both safety and operational efficiency. The TEPIDOIL framework can now be expanded to include:

- Integration
- Safety
- Climate Change and Sustainability
- Capability Protection and Security
- Regulation
- TEVV
- Third Party Capacity
- Political, Sociological and Ethical
- Commercial
- Financial

Also, by combining TEPIDOIL and the additional categories with NIST's Risk Management Framework these wider considerations can be included into a full spectrum of risk assessment using a 3D risk assessment model.

The Artificial Intelligence Risk Management Framework (AI RMF) is designed to manage AI risks and ensure the responsible development of trustworthy AI systems. It consists of four core functions: GOVERN, MAP, MEASURE, and MANAGE, each with specific categories and actions (NIST 2023).

Govern function cultivates a risk-aware culture in an organisation to ensure responsible AI system development. It involves risk anticipation, identification, and management, aligning with organisational principles, and enhancing competencies of AI system personnel. It also deals with third-party software, hardware, and data issues.

Map function provides context for understanding LAWS risks. It considers AI lifecycle complexities and their impacts on operational environments. By anticipating and addressing risks, it enhances decision-making.

Measure function uses varied methods to analyse and monitor AI risks and their impacts. It includes testing LAWS systems before and during operation. Risk measurements cover system functionality, trustworthiness, social impact, and human-AI interactions, providing a basis for management decisions and improving testing.

Manage function allocates resources to mapped and measured risks. It includes risk treatment plans for incidents and gathers input from stakeholders to reduce system failures and negative impacts. It ensures ongoing risk management, transparency, and accountability.

The Safety Assurance of Autonomous Systems in Complex Environments (SACE) methodology, developed by researchers at the University of York (R. Hawkins et al. 2022), aligns with UL4000, UL4600, and SCSC-153B standards for autonomous road vehicles. However, it lacks coverage of military applications and safety deviations authorised by duty holders.

The SACE process runs parallel to typical systems engineering assurance and safety case activities, assuming that a conventional safety case is already established before incorporating an autonomous safety case. SACE outlines specific characteristics for an autonomous safety case:

- Safety requirements are identified at multiple levels (System, Subsystem, Component).
- Requirements are traceable and maintain system integrity.
- Requirements can be allocated to implementing components.
- Verification at each level provides evidence of requirement satisfaction.
- Constant analysis detects common cause failures reflected in requirements.

SACE is designed for new systems and works by tracing through each subsystem of the autonomous system. Verification at the subsystem level does not include individual component requirement sets, as specific components, like machine learning (ML), have their own guidance (e.g. Assurance of Machine Learning for Autonomous Systems – AMLAS) (R. Hawkins et al. 2021).

However, SACE has limitations as it focuses on a single autonomous system and subsystems and is highly tailored, ignoring human factors. LAWS potentially operating with deviations under Duty Holder control must consider interactions with humans, traditional machines, and other autonomous systems within a broader system of systems. Therefore, this and other recent frameworks addressing AI risks are considered insufficient for LAWS challenges (Janjeva et al. 2023). The Centre for Emerging Technology and Security (CETAS) paper (Janjeva et al. 2023) for example operates similarly to SACE; however, it emphasises addressing “harms” rather than “risks,” focusing on community experiences in the presence of an autonomous system, rather than potential community risks.

The AI RMF and the SACE frameworks emphasise the importance of risk management, traceability of requirements, and continuous monitoring of AI and autonomous systems. However, there are notable differences. AI RMF provides a structured approach with core functions and highlights the multidisciplinary aspect, considering external actors. SACE, on the other hand, focuses on specific characteristics for autonomous safety cases and may have limitations when dealing with LAWS operating under Duty Holder control and interactions in complex environments. The AI RMF seems to have a more comprehensive and adaptable approach, while SACE is tailored specifically for civilian autonomous systems and may need further adaptation to address LAWS operating within a system of systems. Other models such as the OECD AI Principles and EU AI

Act were ruled out as they are more specific and legally binding, focussing on regulation and compliance, which could potentially slow down AI innovation.

The Council of Europe's AI and Human Rights paper and UNESCO recommendations are very ethically oriented, emphasising fundamental human rights and values. However, they lack the force of law. The EU AI Act was observed to be the most legally binding and comprehensive framework but it may face challenges in achieving the right balance between regulation and innovation.

All these frameworks highlight the importance of ethics, human rights, and democratic values in AI development, reflecting a growing global consensus on these issues. However, the current lack of a uniform global regulatory framework for AI remains a challenge as it can lead to discrepancies in AI governance and ethical standards across regions.

In summary assurance in LAWS is crucial to ensure their safe and reliable operation. Traditional industry assurance relies on TEVV during development, but autonomous systems require continuous runtime assurance due to their capacity to respond unpredictably to unforeseen situations. This need for ongoing assurance incurs significant temporal and financial costs, driven by the necessity for extensive evidence supporting safety-critical decisions. The assurance process must ensure adherence to ethical values such as transparency, fairness, non-maleficence, responsibility, and privacy, which are essential for maintaining trust and reliability in AI systems. As Leslie (2019) highlights, AI assurance begins with process-based and product-based assurance, yet autonomous systems demand additional tailored principles due to their diverse applications.

Collaboration between engineers, scientists, and domain experts is critical in developing governing rules for autonomous system behaviour. Machine learning is a pivotal component in LAWS, enabling AI systems to learn, adapt, and improve from data. Ensuring the quality and integrity of training data is paramount to prevent unintentional biases, which can arise from sampling, labelling, or measurement errors. Such biases can perpetuate and amplify through feedback loops, significantly impacting the system's performance and fairness. Rigorous data validation, third-party oversight, and the use of independent or synthetic data resembling the operational environment are necessary to mitigate these risks. Additionally, effective risk mitigation strategies, such as multi-layered security approaches, real-time monitoring, and robust incident response plans, are crucial to safeguard LAWS in complex and adversarial environments, ensuring they operate safely and effectively even under challenging conditions (DSTL (2021); Li et al. 2023).

5. Operating itself safely

A review of the literature has established that the merging of "safe to operate" and "operate safely" is not only feasible but vital for the assurance of autonomous systems. The "Operating Itself Safely" method exemplifies this convergence. This section presents an "Operating Itself Safely" theoretical assurance framework for use with systems that are able to seed control over to AI.

An example of such a system could be a fighting vehicle that can be driven manually or operate semi-autonomously with aids and assists for navigation, driving, and weapons systems. If this vehicle were to have a fully autonomous mode where full control of some

or all systems is ceded to an AI, it could be deployed uncrewed to scout a minefield, for example. In this scenario, the “Operating Itself Safely” concept would be enacted, ensuring that the AI manages the vehicle’s operations effectively and safely without human intervention.

There is a hierarchy of documents detailing safety rules within the Royal Navy (these documents also have wider relevance to the entire MOD). As a result, the overall safety culture is a product of the general military ethos and culture. The Navy promotes an ethos based on “Safe to Ourselves and Dangerous to the Enemy” through the acronym C2DRIL – Courage, Commitment, Discipline, Respect, Integrity, and Loyalty (Royal Navy 2016). This culture works well for a military organisation but does not always lend itself well to an industry that is heavily involved in technology and safety. The safe operation of technology must always be paramount in peace time, however in operational circumstances other factors are often prioritised above health and safety. In November 2022, ChatGPT, a Large Language Model (LLM), was launched (Lock 2022). Inspired by this, Palantir, a software company, revealed plans for a military Artificial Intelligence Platform (AIP) with LLM-like capabilities (Palantir 2023). The AIP’s goals include integration with military systems, including mission-critical ones. In promotional materials (Bowman 2021), Palantir’s system demonstrated features including enemy movement alerts, autonomous drone deployment, and engagement options. This advanced technology raises concerns about automation bias,² where users must decide on system recommendations or seek human validation for potentially lethal actions. Over time, there is a risk of users blindly trusting the system while retaining full responsibility within the kill chain. A case in point is the US Defence Advanced Research Projects Agency (DARPA) project “Urban Reconnaissance through Supervised Autonomy” (URSA) that initially led to hasty conclusions from operators when it categorised individuals as “threats,” marked as red dots (Michel 2023). To address this issue, the system was directed not to label humans as targets and instead provide information about “persons of interest.”

Drawing from these lessons, Palantir is addressing this problem of bias by deliberately introducing inefficiencies into automated processes with methods such as asking for unnecessary approvals or mandating operators to confirm recommendations using a second independent source of intelligence (Bowman 2021).

Given the lessons learned and the imperative to include controlled inefficiencies, forthcoming assurance models for AI technologies must proactively counteract automation, assimilation, and cognitive biases. Protecting human judgment against the undue influence of highly intelligent machine systems can be accomplished through subtle design elements that avoid unintentionally swaying human decision-making, such as neutral graphical representations (Crotoft 2022). It is crucial to acknowledge that introducing inefficiencies into systems to mitigate biases will impact the system’s use cases and requirements.

Live military operations are inherently unpredictable, making it difficult for commanders and operators to anticipate adversaries’ weapon usage and plan their responses. Introducing AI and autonomous features further complicates this unpredictability. Predictability, essential for ensuring system accuracy, transparency, explainability, and interoperability, encompasses both technical and operational aspects. Technically, it reflects a system’s ability to perform tasks reliably, as demonstrated during testing.

Operationally, it considers the predictability of an autonomous system's actions, influenced by factors like environmental complexity, adversary interactions, and operator understanding. LAWS may behave unpredictably when facing unforeseen inputs or operating beyond their intended environment.

Predictability also applies to human-machine interaction in an environment. Delegating tasks to LAWS is unpredictable due to the inability to account for every possible scenario. This unpredictability raises concerns about meaningful human control (Bronk, Bruun, Connolly, Davison, Gettinger, Hinds, Huelss, Kayser, Mohan, Nadibaidze, and Moyes 2023). International consensus deems the risk of LAWS using force without human assessment unacceptable. The solution lies in integrating human control into the design of LAWS and their intended use.

Ensuring the safe operation of autonomous systems presents challenges; Predictable inputs may not always result in predictable outputs. For example In an urban combat environment, a Lethal Autonomous Weapon System (LAWS) may receive a visual input showing a person holding a firearm, which it predictably classifies as a hostile target. However, this can lead to unpredictable outputs if the individual is actually an ally or civilian. Additionally, adversaries might exploit this by creating decoys, leading to false positives, or environmental changes such as smoke might cause misclassification of objects and individuals (Cummings 2017). Systems that operate without operator input require continuous training and monitoring to maintain lifelong safety. Dynamic environmental factors can push a system beyond its intended operating context, while static safety cases prove inadequate for machine learning models gaining experience. Traditional assurance and audit techniques become obsolete post-completion, thereby shifting trust towards the system's training data and internal validation processes, thus preserving some aspects of trustworthiness through assurance. These challenges intersect with Figure 4, illustrating the components of the current Royal Navy assurance model. Traditional artifacts, marked in orange, are no longer suitable for assuring autonomous systems, so a new dynamic approach is demanded. In the case of an autonomous system being "safe to operate itself safely" the safety systems and policies highlighted in orange will fall short and require review.

At the 2019 CCW conference, discussions centred on challenges related to quantifiability, context, and predictability of system output. Two contrasting viewpoints emerged: one seeking technological solutions, the other viewing these issues as inherent limitations in technology. The consensus was that human control and autonomy must coexist (Boulanin, Davidson, Goussac, and Carlsson 2020). Advocates emphasised three key objectives for LAWS deployment:

- Control over usage parameters.
- Control over the operational environment.
- Control through human-machine interaction.

To ensure compliance with IHL, ongoing human supervision is necessary. Relying solely on control-by-design is insufficient as continuous spatial and temporal assessments are vital to manage the system's unpredictability. In complex scenarios, like a busy battlefield or shipping lane, these assessments may become challenging, potentially pushing the

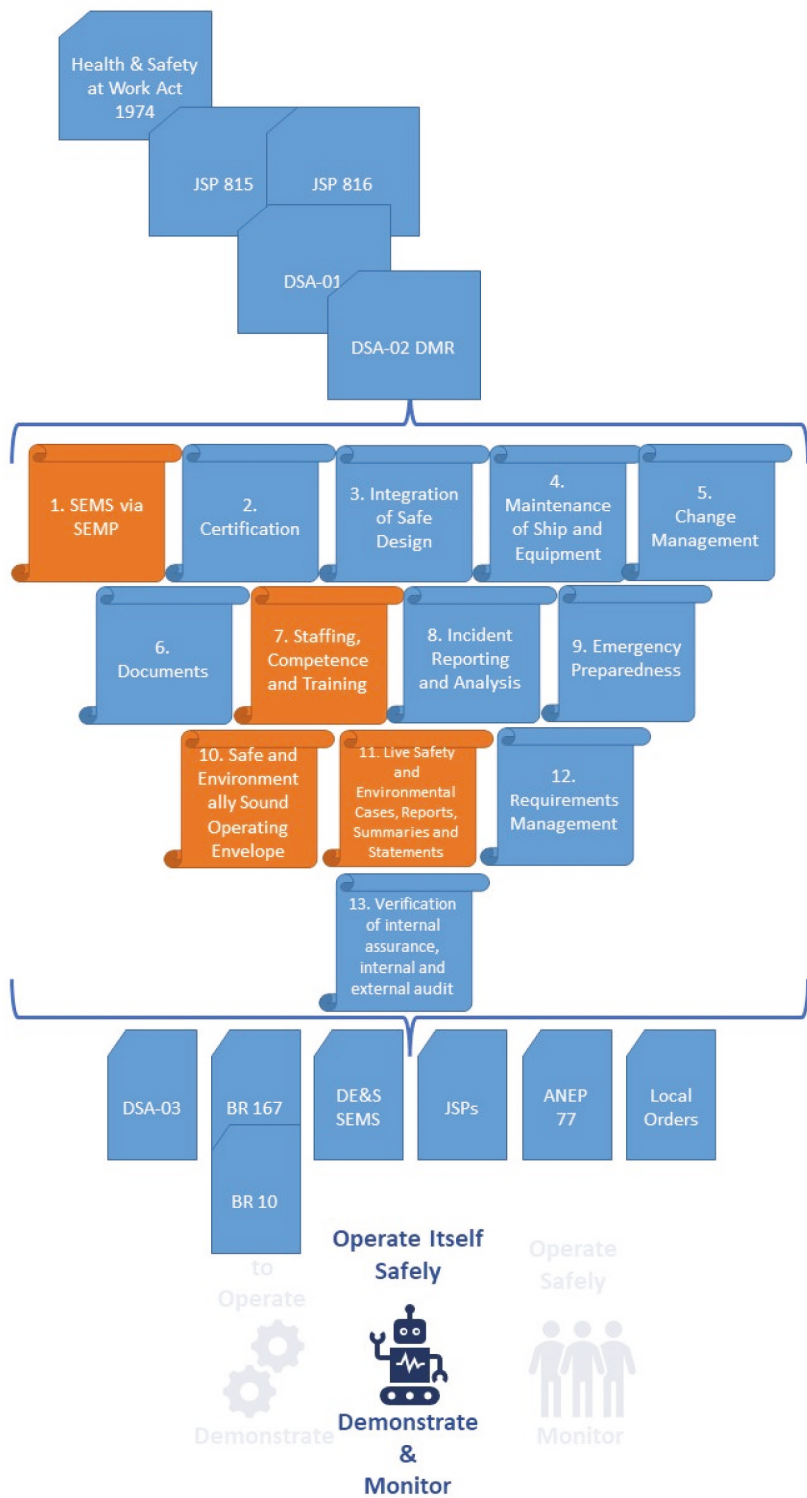


Figure 4. Safety hierarchy – operate itself safely. Spayne, 2023.

system beyond its defined scope. In operational scenarios where ground conditions change post-LAWS deployment, rendering planning estimates obsolete, operators must retain the ability to intervene or deactivate the system for IHL compliance. Challenges arise in certain domains, for example sub-surface operations, where reliable communications are not guaranteed. As such, ensuring “meaningful human control” by a “responsible agent” is essential but not always possible in all but the most benign environments. This raises ethical questions about operational capabilities and constraints, as rules on targeting are fundamentally rooted in probability-based assessments and moral reasoning.

IHL rules depend on the situation, and some argue that humans do not have to make all decisions because these rules intentionally have some uncertainty to avoid technical arguments. Proportionality, which depends on personal judgment, can be seen as either reflecting real-world complexities or suggesting that deciding on targets should also consider individual viewpoints. The lack of consensus in current doctrine makes it challenging to argue that rules around the conduct of hostilities explicitly mandate human judgment and control. However, it is essential that the design and use of LAWS provide operators with the capability to make judgment-based decisions (Sassòli 2014). Therefore, deploying LAWS without human supervision would be considered unlawful because IHL’s targeting rules require computation techniques to be on par with human capabilities (Horowitz 2016), noting that this is contingent on current realities and technological constraints.

The CCW meetings aimed to establish new rules and regulations that would anticipate potential developments in LAWS technology. They proposed that these rules should be broad enough to accommodate emerging technologies that might eventually replicate human reasoning (Rešlová and Sassòli 2022). These rules would entail programming the LAWS with algorithms that guide their decision-making processes, similar to human decision-making.

When autonomous systems enter combat independently, it will mark a significant shift from humans using tools for war to LAWS conducting wars independently. There are no specific rules against this yet, as existing IHL rules mainly focus on action outcomes and do not cover decision-making processes. Nevertheless, human control is involved in development, especially in designing weapons. Some argue that human judgment should be implicit in every lethal force decision during the authorisation process (Asaro 2012).

5.1. Assuring LAWS within the royal Navy’s duty holder construct

The operating principles governing operational environments establishes a clear division of responsibilities between the duty holder and the commander. The duty holder is accountable for ensuring that deployed forces are well-versed in high-risk activities within a specified operating framework (Ministry of Defence 2022). The duty holder maintains responsibility over the deployed forces as long as they remain within this established framework. If a particular situation necessitates taking greater risks beyond what was initially planned, the CO can order such actions, provided they can justify the benefits. In such cases, the CO assumes full responsibility for the added risk (Defence Maritime Regulator 2023).

Front Line Commands typically bear the majority of risks during live operations. Supporting these risks are duty-holder-facing-organisations that provide essential services and infrastructure. Often, it is individuals within these organisations who effectively *manage* RtL indirectly through the engineering and logistics services they offer.

When considering LAWS and autonomous systems that straddle the categories of safe to operate and operating safely, they can be viewed as infrastructure tools in addition to being viewed as standalone capabilities. Therefore, individuals responsible for software algorithms are just as vital to the kill chain as the commanders or soldiers implementing their use on the ground. As a result, engineering decisions behind the scenes could potentially push a system beyond its operating framework even before deployment, necessitating increased assurance or inclusion in a combat risk assessment.

The current military structure follows a long-standing linear and hierarchical tradition but embracing groundbreaking technologies like LAWS requires a rethink of decision-making processes (Balis and O'Neill 2022). To effectively integrate LAWS, it is crucial to reimagine the role of humans in warfare. Training stakeholders to understand and interact with these systems is a vital first step to harness the synergy between inputs, AI, and LAWS' outputs and establish confidence at an organisational level. Future commanders need to master the deployment of assets that are not merely under human control but are programmed with intent and the ability to devise their methods. The Atlantic Council emphasises the transformative potential of autonomous systems in military operations. Autonomous systems can handle a wide range of tasks, from data analysis to real-time asset management, freeing military personnel to focus on strategic decisions. This integration requires commanders to understand and trust these systems to act autonomously, ensuring efficient and effective mission execution (Atlantic Council 2022). This necessitates a reassessment of human roles within a broader human-machine cognitive system.

Current technical challenges impede LAWS from operating autonomously at the standards required to uphold safety and reliability in deployed operations under existing regulations. Despite recent technological progress, LAWS still fall short of the necessary standard to function effectively in complex and unpredictable environments. This is mainly because perception and decision-making capabilities remain at a low Technology Readiness Level. While machine vision may excel at identifying static images and labelling elements in moving video, even in real-time, it struggles with discerning intent and context (Edell 2023) – a superpower that the human brain excels at, however. In practice, this means that these systems can be easily deceived, as demonstrated by a test conducted by US Marines against an AI-based security system. They outwitted the system without firing a single shot, simply by concealing themselves in a box and nonchalantly crawling past it (Bandara 2023).

This highlights that while LAWS can effectively respond to numerous anticipated scenarios within a controlled operating environment, they remain vulnerable to unforeseen tactics or techniques not accounted for in their initial design or subsequent machine-learning adaptations. Anything beyond the system's planned capabilities or unaddressed during its lifecycle poses a significant vulnerability. The inherent unpredictability associated with autonomy is likely to continue posing challenges to operational

safety, irrespective of the system's capabilities (Boulanin, Davidson, Goussac, and Carlsson 2020).

To ensure human agents' safety in most foreseeable scenarios, human attendants must be retained to oversee LAWS and be ready to intervene in edge cases where the system cannot navigate challenges effectively. Sharkey (2016) stressed the importance of human control in aligning LAWS with the operation's overall strategy, adhering to local rules of engagement, and leveraging humans' cognitive capabilities, including contextual understanding, tactical and strategic insight, qualitative judgment, and reasoning. This collaborative human-machine approach optimally combines the machines' rapid target acquisition and precision with humans' contextual appreciation and judgment.

The literature examined in this paper has discussed approaches employed in civilian industries and methods developed for general autonomous systems. It has also highlighted deficiencies in the existing assurance techniques utilised by the Royal Navy and briefly addressed ethical, cultural, and traditional concerns. To address such

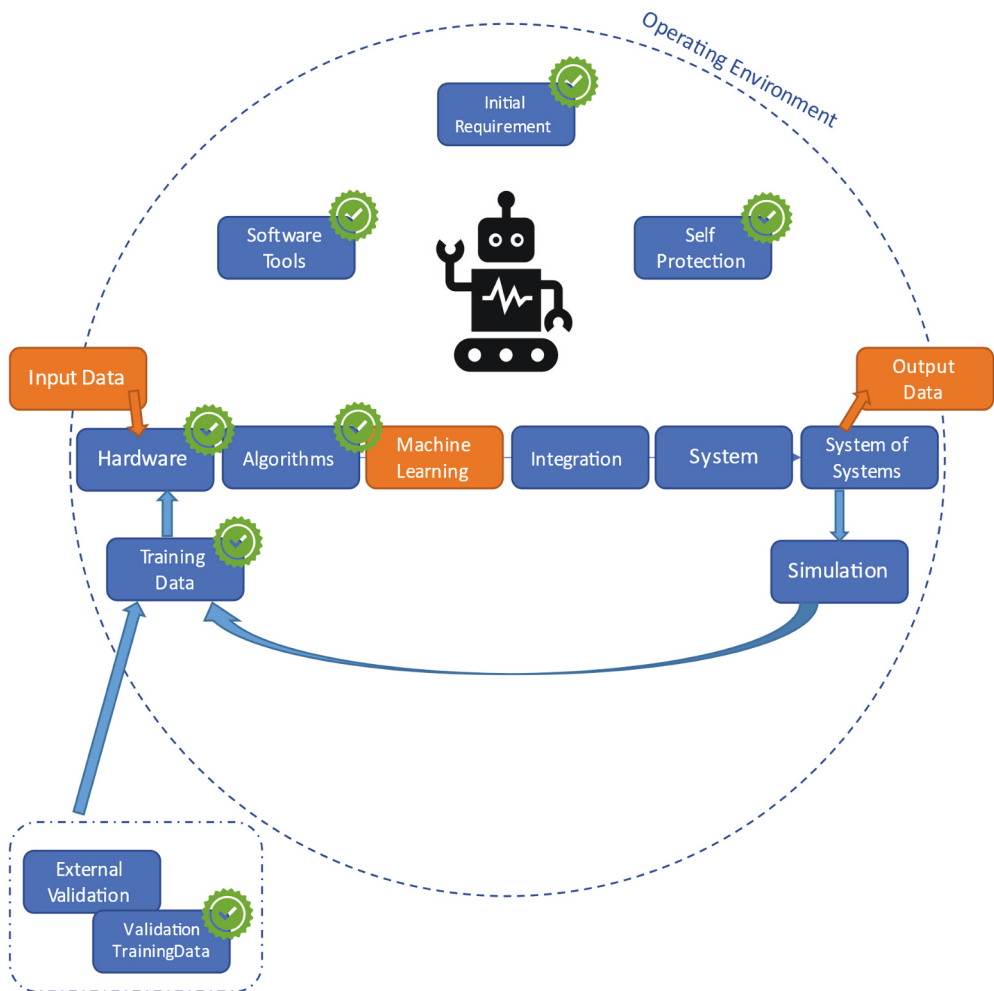


Figure 5. Operating itself safely. Spayne, 2023.

deficiencies, a methodology taking aspects of “safe to operate” and “operate safely” can be formulated into an initial theory for assuring LAWS, where aspects for individual assessment are depicted (Figure 5). Figure 5 depicts the individual areas of an autonomous system within the operating itself safely framework that all require individual attention. It is recommended that in a safety case each label would be its a standalone chapter. The image depicts elements that interlink in sequence and elements that sit in and outside of the systems boundary. Noting that an attempt to control the operating environment in the same way that the training environment would be manipulated is a valid way to ensure safety. (For example, the area where a LAWS might be deployed could be labelled an exclusion zone with civilian actors forbidden to enter). Blue boxes indicate aspects that may be assured traditionally via equipment authorities hoping to achieve “safe to operate.” Orange boxes indicate unpredictable and unmappable aspects that would traditionally have been captured via training and assessment for Operate Safely were the system being controlled manually by a human agent.

The reviewed literature suggests that traditional safety cases and risk analysis techniques should be utilised when conventional technology and methodologies are present within the system of systems. This approach is to be integrated when a system has self-operating capabilities. The entire system should be assured within the context of its operation in a specific operational environment, under the supervision of a well-equipped organisation. Even then, the system can only receive *interim* accreditation as being safe to “operate itself safely” before deployment. Only after a substantial period of live operational use, during which reliability is consistently demonstrated, documented, and visually verified, with expected outputs routinely observed in a live environment, can a system be deemed *fully* “safe to operate itself safely,” noting there will still be limitations.

Once deployed, the AI-based system is treated, from an assurance perspective, akin to a human. Like humans, whose education (simulated input data) involves training and live experiences, AI systems follow a similar trajectory. Their output data (actions) is continuously assessed based on ongoing performance and responses to unexpected inputs. Both humans and AI systems share the uncertainty of potential unexpected behaviour due to unforeseen circumstances. Despite the inherent risk of unexpected responses, repeated satisfactory performance allows humans to establish a reputation for reliability. In the case of LAWS, while they can never be fully guaranteed to be safe, their safety can be considered on par with human reliability.

Therefore, after the hardware and non-AI software within the systems are assured against their requirements, in a defined operating area, under the control of a well-resourced organisation, the remaining assurance activities must focus on assuring the training input data and monitoring the live outputs in response to live environmental inputs for expected results. The assurance model, visualised as an “onion” illustrates the system of interest, its subcomponents, and its place within a broader organisation and wider operating area (Figure 6). The layers of this model are explained below to depict the multiple levels at which assurance activities must take place.

Layer -2 (Computation): Identify and assure all computations, with a focus on safety-critical functions where reasonably practicable.

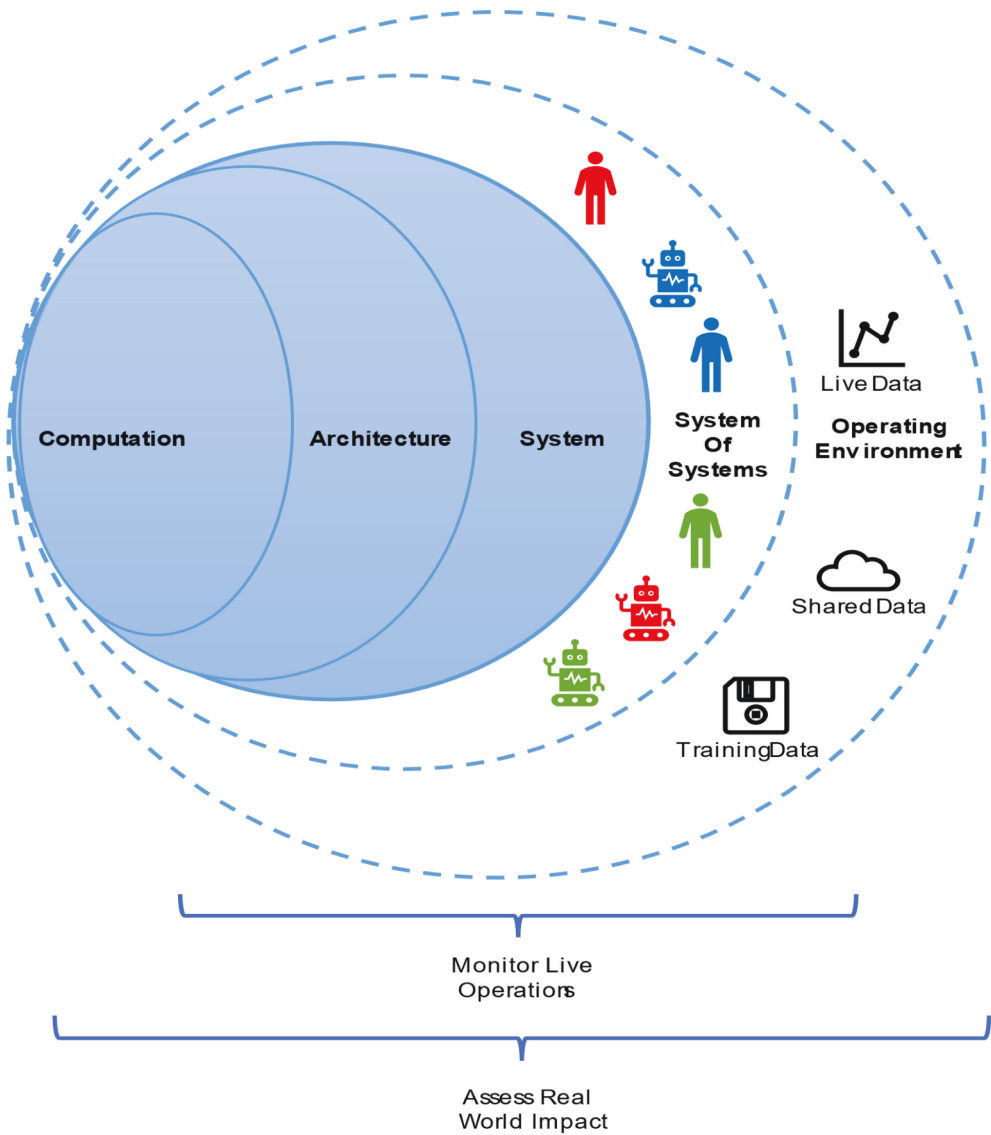


Figure 6. LAWS assurance onion. Spayne, 2023.

Layer -1 (Architecture): Define the system architecture, assure traditional hardware and software using established methods, and appropriately assure machine learning and autonomous models following this methodology.

Layer 0 (System): Establish system boundaries, set requirements for operators and interacting agents to ensure they operate and interact with the system safely, conduct risk analysis for traditional system aspects to ensure traditional hardware and software components are safe to operate.

Layer 1 (System of Systems): Identify the intended and unintended nodes in the wider system of systems, including other autonomous systems, traditional systems, and humans that may interact with this platform within the operational environment. Document likely

behaviours and conduct risk analysis based on the roles of other agents and nodes as operators, friendly entities, neutral bystanders, or adversaries.

Layer 2 (Operating Environment): Define the operating environment, identify sources of live data, potential sources of shared live data, and actively feed this back to training data creators for curation and reissue, to optimise the system for operations in a specific environment.

Layer 2.1 (Data): Assure and validate training data, assure and validate secondary sources and third-party (shared) validation data, and ensure data validation and curation tools are reliable. Also, assure the integrity of the simulated training environment.

Layer 3 (Monitoring Live Operations – Deployment Phase): Implement live monitoring of operations, analyse the returned data, and update assurance activities from Layer -2 upwards based on continuous feedback.

Layer 4 (Real World Impact): Evaluate the system's impact on real-world communities, considering social, political, and economic influences in the operating environment.

5.2. Conducting assurance

In conclusion to the research and analysis throughout this paper, the proposed assurance process involves six steps. Each step must consider its impact on each layer and the operating environment to assure the safe operation of LAWS.

Step 1 prepares the receiving organisation for LAWS procurement by setting the highest-level capability requirements, terminology, and definitions, involving stakeholder engagement across the TEPIDOIL+ spectrum to outline required governance before the typical engineering procurement cycle. The U.S. Department of Defence's AI Adoption Strategy emphasises the importance of establishing clear capability requirements and terminology to guide AI integration into military operations, highlighting the need for stakeholder engagement across various domains to ensure comprehensive governance (Defence Department News, 2023). Additionally, a RAND Corporation report discusses the necessity of setting clear definitions and high-level requirements to streamline the procurement and deployment of advanced military technologies, ensuring alignment with strategic goals and operational needs. (Martin et al. 2019)

Step 2 focuses on cartography, where requirements and broader considerations are mapped for the entire system of systems, taking into account the intended operating environment and context. The integration of the Joint All-Domain Command and Control (JADC2) in the U.S. military serves as an example, illustrating how requirements and considerations are mapped across a complex system of systems while accounting for diverse operational environments. Empirical evidence from the NIST AI Risk Management Framework highlights the importance of mapping AI system requirements within their intended operational contexts to ensure comprehensive risk assessment and management. (NIST 2023)

Step 3 conducts a global risk assessment and analysis, emphasizing the upgrade of traditional two-dimensional risk assessments to a custom three-dimensional Autonomous risk matrix based on the ASIL (Automotive Safety Integrity Level) system used in the automotive industry. The ASIL system, which is adapted for assessing risks in autonomous vehicles, serves as a model for this enhanced risk assessment approach (ISO

26,262). Research by the University of York's Safety-Critical Systems Research Group demonstrates the application of ASIL principles in evaluating the safety and reliability of autonomous systems across various domains, highlighting the effectiveness of this three-dimensional risk assessment methodology. (R. D. Hawkins and Kelly 2009)

Step 4 involves addressing the data collected in Steps 1-3 through a series of Go/No-Go decision-making questions to determine the costs and feasibility of continuing the project, also considering whether the system actually needs to be autonomous. The decision-making framework used in NASA's project management includes Go/No-Go checkpoints to evaluate project feasibility, costs, and the necessity for autonomous capabilities, ensuring only viable projects proceed (NASA 2020). Similarly, the DARPA utilises decision-making processes to rigorously evaluate the feasibility and strategic value of new technology projects before further investment, ensuring thorough assessment and resource allocation. (DARPA 2023)

Step 5 involves active monitoring during system development, model training, and deployment. Continuous monitoring systems used in autonomous vehicle development, such as Tesla's Autopilot, involve real-time data analysis during development and deployment to ensure safety and performance (TESLA 2021). Studies by the MIT Lincoln Laboratory emphasise the importance of active monitoring and feedback loops in AI system development to adapt to new data and operational environments, ensuring ongoing safety and reliability. (MIT Lincoln Library 2020)

Step 6 outlines a disposal and emergency withdrawal plan that adheres to ethical guidelines and considers data security aspects. The decommissioning process of autonomous systems in the defence sector includes detailed protocols for data sanitization and secure disposal, ensuring compliance with ethical guidelines and data security standards (UK MOD 2022). Research published in the Journal of Cybersecurity discusses the critical need for secure data disposal practices in autonomous systems to prevent unauthorized access and ensure that valuable data can be repurposed for future training. (Alston and Bird 2023)

The proposed assurance process for LAWS involves a comprehensive approach that considers the impact and influence of each layer and the operating environment within every step. Each step – from preparation and cartography to risk assessment, decision-making, monitoring, and disposal – is crucial for ensuring the safe and effective operation of LAWS. Upon initial deployment and completion of Step 4, the system receives interim accreditation as “safe to operate itself safely.” Step 5 then begins with a feedback loop to Step 2, enabling the system to work towards full accreditation based on consistent safe performance. This accreditation may be withdrawn if an unexpected event requires investigation, at which point the system will either be modified for reaccreditation or proceed to Step 6 for disposal and data analysis. Disposal at Step 6 involves complex procedures, including data sanitisation for security and recycling into training data stock for future systems.

6. Conclusion

In this study a multifaceted approach to assuring the safety of autonomous systems has been explored, with a particular focus on the emerging concept of “Operating Itself

Safely.” Traditional assurance methods, such as TEVV, have served as the bedrock of safety assurance in various industries. In the military, the duty holder construct adds a further layer of complication to ensuring safety in situations where the usual regulations do not apply. The unique characteristics of autonomous systems and LAWS demonstrate areas where traditional methods are no longer fit for purpose. Therefore, attempting to resolve this problem demands a paradigm shift in safety assurance.

An assurance process must not only build confidence but also contend with the inherent unpredictability of autonomous systems and their capacity to adapt to unforeseen situations. The novel framework proposed in this research achieves this in theory by addressing intricate aspects of AI assurance, emphasising process and product-based assurance, and drawing attention to compliance with both human and AI ethical values. The inherent complexity of LAWS, often operating in uncertain and non-human-perceptible environments, necessitates software based ongoing runtime assurance to address unexpected scenarios effectively. Hardware reliability, operational speed, and the importance of real environmental data have been underscored as vital considerations in assuring autonomous systems’ safety.

A detailed literature review and data collected via interview has shed light on the significance of training data, avoiding unintended biases, and the diversity of algorithms and their respective assurance requirements, that ensure algorithm reliability, specific system-level requirements, and balanced synthetic and real-world testing. All are crucial aspects of assuring LAWS. With equal importance is the interplay between various systems and the interactions with humans, adding another layer of complexity to the assurance process.

The wider implications of this research expand into all industries that may wish to utilise autonomous systems. The Royal Navy was chosen for this case study as the uniqueness of the duty holder construct, military operations and the ability to encompass weapon systems allow for a greater range of considerations to be factored. Further research shall take place to continue developing and optimising the proposed assurance model to assess a systems competence to “operating itself safely.”

Notes

1. A self-contained weapon system with the capability to acquire, discriminate, classify and prosecute a target as governed by IHL without human intervention.
2. Automation Bias. When one follows the recommendations of an AI system blindly without verification. For example: Blindly following a SATNAVs direction into a field despite knowing a better route.

Glossary

AAIP	Assuring Autonomy International Program
ACOPS	Approved Codes of Practice
AI	Artificial Intelligence
AI HLEG	Artificial Intelligence High Level Expert Group
ALARP	As Low As Reasonably Practicable
ALTA	Assessment List for Trustworthy AI
AMLAS	Assurance of Machine Learning for Autonomous Systems

ANEP	Allied Naval Engineering Publication
ARP	Aerospace Standards Protocol
ASIL	Automotive Safety Integrity Level
BR	Book of Reference
C2DRIL	Courage, Commitment, Discipline, Respect, Integrity, Loyalty
CCA	Common Cause Analysis
CCW	Committee on Conventional Weapons
CMA	Common Mode Analysis
CO	Commanding Officer
DAL	Development Assurance Level
DARPA	Defence Advanced Research Projects Agency
DCavg	Diagnostic System Convergence
DE&S	Defence Equipment and Systems
DEFSTAN	DEFence STANDard
DMR	Defence Maritime Regulator
DSA	Defence Safety Agency
DSTL	Defence Science and Technology Laboratory
EU	European Union
FHA	Functional Hazard Analysis
FMEA	Failure Mode and Effect Analysis
FMES	Failure Mode and Effect Summary
FTA	Fault Tree Analysis
GGE	Group of Government Experts
GPT	Generative Pre-Trained Transformer
HASAWA	Health and Safety at Work Act
HSE	Health and Safety Executive
ICRC	International Committee of the Red Cross
IEC	International Engineering Council
IHL	International Humanitarian Law
IMO	International Maritime Organisation
ISO	International Standards Organisation
JSP	Joint Services Publication
LAWS	Lethal Autonomous Weapon System
LLM	Large Language Model
MARPOL	Maritime Pollution
ML	Machine Learning
MOD	Ministry of Defence
MTTFd	Mean Time To Failure
OC	Operational Capability
OECD	Office of Economic and Cooperative Development
PFHd	Probability of Failure per Hour
PL	Performance Level
PRA	Particular Risk Analysis
PSSA	Preliminary Safety System Analysis
QM	Quality Managed
RMF	Risk Management Framework
RN	Royal Navy
RtL	Risk to Life
SACE	Safety of Autonomous Systems in Complex Environments
SCSC	Safety Critical Systems Club
SDA	Submarine Delivery Agency
SEMP	Safety and Environment Policy
SEMS	Safety and Environment System
SFHA	System Functional Hazard Analysis

SIL	Software Integration Level
SOLAS	Safety Of Life At Sea
SOUP	Software Of Unknown Provenance
SSA	System Safety Assessment
TEPIDOIL	Training, Equipment, Personnel, Information, Doctrine, Organisation Infrastructure, Logistics
TEVV	Testing, Evaluation, Verification, Validation
TRL	Technology Readiness Level
UK	United Kingdom
UL	Underwriters Laboratories
UNESCO	United Nations Education Science and Cultural Organisation
URSA	Urban Reconnaissance Through Supervised Autonomy
ZSA	Zonal Safety Analysis

Disclosure statement

No potential conflict of interest was reported by the author(s).

Notes on contributors

Peter Spayne is a Weapon Engineer in the RN, majoring on mine disposal systems. He studied Autonomous systems at BEng and explored weaponising high powered ultrasound to agitate submerged explosives at MSc with Staffs Uni. He is a PhD student researching the assurance of lethal autonomous weapon systems at Cranfield University.

Dr Laura Lacey is a Senior Lecturer in Military Aviation Safety and Airworthiness at Cranfield University.

Dr Marie Cahillane is Head of the Applied Psychology Group and a Senior Lecturer in Applied Cognitive Psychology within the Centre for Electronic Warfare, Information and Cyber at Cranfield University. She has over 15 years' experience in leading and collaborating on defence and Security research.

Prof Alistair Saddington is Professor of Defence Aeronautics. He has over 25 years of experience in aerospace engineering across both industry and academia including research leadership, management, and postgraduate teaching and supervision.

References

- Alston & Bird. 2023. *Secure Data Disposal and Data Minimization*. <https://www.alston.com/en/insights/publications/2023/02/secure-data-disposal-and-data-minimization>.
- "Article 49 - Definition of Attacks and Scope of Application. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts (Protocol I)." 1987. <https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-49?activeTab=1949GCs-APs-and-commentaries?>.
- 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. <https://doi.org/10.1017/S1816383112000768>.
- Atlantic Council. 2022. "What Does the Future of Autonomous Warfare Look Like - Four Critical Questions Answered." *Automating the Fight*. <https://www.atlanticcouncil.org/content-series/automating-the-fight/what-does-the-future-of-autonomous-warfare-look-like-four-critical-questions-answered/>.

- Balis, C., and P. O'Neill. 2022. "Trust in AI: Rethinking Future Command." <https://www.rusi.org/explore-our-research/publications/occasional-papers/trust-ai-rethinking-future-command>.
- Bandara, P. 2023. *U.S. Marines Outsmart AI Security Cameras by Hiding in a Cardboard Box*. PetaPixel Inc. <https://petapixel.com/2023/01/30/u-s-marines-outsmart-ai-security-cameras-by-hiding-in-a-cardboard-box/>.
- Bills, G. 2014. "LAWS Unto Themselves: Controlling the Development and Use of Lethal Autonomous Weapons Systems." *Geo Wash L Rev* 83:176.
- Booker, C., and J. Wright (Director). 2016. *Black Mirror: Nose Dive*. Netflix.
- Boothby. 2019. "The Law of Targeting." *International Humanitarian Law: Rules, Contravencies and Solutions to Problems Arising in Warfare*.
- Boulanin, V., and M. Verbruggen. 2015. *Mapping the Development of Autonomy in Weapon Systems*. SIPRI.
- Boulanin, Davidson, Goussac, & Carlsson. 2020. *Limits on Autonomy in Weapon Systems*. www.sipri.org.
- Bowman. 2021. *Palantir's Response to NTIA on AI Accountability Policy*. Palantir Blog. <https://blog.palantir.com/palantirs-response-to-ntia-on-ai-accountability-policy-428ebb33698d>.
- Bronk, Bruun, Connolly, Davison, Gettinger, Hinds, Huelss, Kayser, Mohan, Nadibaidze, & Moyes. 2023. "Loitering Munitions and Unpredictability Autonomy in Weapon Systems and Challenges to Human Control." <https://doi.org/10.5281/zenodo.7860762>.
- Conn, A. 2017. "Future of Life Institute." *Beneficial AI*. <https://futureoflife.org/event/bai-2017/>.
- Crootof, R. 2022. In *The Ethics of Automated Warfare and Artificial Intelligence*. Centre for International Governance and Innovation.
- Cummings, M. 2017. *Artificial Intelligence and the Future of Warfare*.
- DARPA. 2023. *Developing Trustworthy AI to Inform Decisions When Every Moment Counts*. <https://www.darpa.mil/news-events/2023-06-06>.
- Defence Department News. 2023. "DOD Releases AI Adoption Strategy." <https://www.defense.gov/News/News-Stories/Article/Article/3578219/dod-releases-ai-adoption-strategy/>.
- Defence Maritime Regulator. 2018. DSA01.2 (Vol. 1).
- Defence Maritime Regulator. 2023. "DSA02 Defence Maritime Regulations for Health, Safety and Environmental Protection."
- DSTL. 2021. *Assurance of Artificial Intelligence and Autonomous Systems - a DSTL Biscuit Book*. <http://www.nationalarchives.gov.uk/doc/open-government-licence/version/3>.
- Edell. 2023. *Recognizing Context is Still Hard in Machine Learning*. Towards Data Science. <https://towardsdatascience.com/recognizing-context-is-still-hard-in-machine-learning-heres-how-to-tackle-it-ed398a725f8c>.
- European Commission. 2022. *Regulatory Framework Proposal on Artificial Intelligence | Shaping Europe's Digital Future*, Shaping Europe's Digital Future. <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>.
- Hall, K. 2010. "Developing Medical Device Software to IEC 62304." *Emdt* -, European Medical Device Technology. <https://www.mddionline.com/software/developing-medical-device-software-iec-62304>.
- Hawkins, R. D., and T. P. Kelly. 2009. "Software Safety Assurance - What is Sufficient?" *4th IET International Conference on Systems Safety 2009 Incorporating the SaRS Annual Conference*: A23–A23. <https://doi.org/10.1049/cp.2009.1542>.
- Hawkins, R., M. Osborne, M. Parsons, M. Nicholson, J. McDermid, and I. Habli. 2022. *Guidance on the Safety Assurance of Autonomous Systems in Complex Environments*. Vol. 1. <https://doi.org/10.48550/arXiv.2208.00853>.
- Hawkins, R., C. Paterson, C. Picardi, Y. Jia, R. Calinescu, I. Habli, and Assuring Autonomy International Programme. 2021. *Guidance on the Assurance of Machine Learning in Autonomous Systems (AMLAS)*. University of York.
- High Level Expert Group on Artificial Intelligence. 2022. "Assessment List for Trustworthy Artificial Intelligence (ALTAI) for Self-Assessment." <https://doi.org/10.2759/791819>.
- Hobbs, & Lee. 2013. *Understanding ISO 26262 ASILs*. Electronic Design. <https://www.electronicdesign.com/technologies/embedded/article/21798166/understanding-iso-26262-asils>.

- Horowitz, M. C. 2016. "Public Opinion and the Politics of the Killer Robots Debate." *Research & Politics* 3 (1). <https://doi.org/10.1177/2053168015627183>.
- Human Rights Watch. 2023. "Background Briefing Review of the 2023 US Policy on Autonomy in Weapons Systems." <https://www.hrw.org/report/2020/10/20/new-weapons->.
- ICRC. 2014. "The Principle of Proportionality." *Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified*.
- International Standards Organisation. 2011. ISO 26262-3: 2011, Road Vehicles — Functional Safety — Part 3: Concept Phase. <https://www.iso.org/obp/ui/#iso:std:iso:26262:-3:ed-1:v1:en>.
- International Standards Organisation. 2015. ISO 13849-1: 2015 - Safety of Machinery — Safety-Related Parts of Control Systems — Part 1: General Principles for Design. <https://www.iso.org/standard/69883.html>.
- Janjeva, A., N. Mulani, R. Powell, J. Whittlestone, and S. Avin. 2023. "Strengthening Resilience to AI Risk."
- Jobin, A., M. Ienca, and E. Vayena. 2019. "The Global Landscape of AI Ethics Guidelines." *Nature Machine Intelligence* 2019 1 (9): 389–399. <https://doi.org/10.1038/s42256-019-0088-2>.
- Kester, L. 2016. "Mapping Autonomy." *Presentation at the CCW Informal Meeting of Experts on Lethal Autonomous Weapon Systems*, Geneva, Geneva.
- Leedeo Academy. 2020. *What Does it Mean That a System is Fail Safe or Intrinsically Safe?*. <https://www.leedeo.es/l/fail-safe-system/>.
- Leslie, D. 2019. "Understanding Artificial Intelligence Ethics and Safety: A Guide for the Responsible Design and Implementation of AI Systems in the Public Sector." *SSRN Electronic Journal*. <https://doi.org/10.5281/ZENODO.3240529>.
- Li, B., P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, and B. Zhou. 2023. "Trustworthy AI: From Principles to Practices." *ACM Comput Surv.* 55(9). 55 (9): 1–46. <https://doi.org/10.1145/3555803>.
- Lock, S. 2022. "What is AI Chatbot Phenomena ChatGPT and Could it Replace Humans." *The Guardian*, December 5.
- Maritime Regulator, Defence 2016. "DSA01.1 Defence Policy for Health, Safety and Environmental Protection."
- Martin, B., D. C. Tarraf, T. C. Whitmore, J. DeWeese, C. Kenney, J. Schmid, and P. DeLuca. 2019. "Advancing Autonomous Systems: An Analysis of Current and Future Technology for Unmanned Maritime Vehicles." *Advancing Autonomous Systems: An Analysis of Current and Future Technology for Unmanned Maritime Vehicles*. <https://doi.org/10.7249/RR2751>.
- Michel, A. H. 2023. "Inside the Messy Ethics of Making War with Machines." *MIT Technology Review*. <https://www.technologyreview.com/2023/08/16/1077386/war-machines/>.
- Ministry of Defence. 2009. *Defence Lines of Development Analysis with MODAF*.
- Ministry of Defence. 2013. *JSP 398, Card Alpha*. 2013th ed.
- Ministry of Defence. 2021. *LAWS Policy Position Statement*.
- Ministry of Defence. 2022. *DSA02-DMR-Defence Maritime Regulations for Health, Safety and Environmental Protection*.
- MIT Lincoln Library. 2020. *Responsible AI Toolbox*. <https://www.ll.mit.edu/r-d/projects/responsible-ai-toolbox>.
- Mobility Insider. 2020. *What is ASIL-D?*. Aptiv.Com. <https://www.aptiv.com/en/insights/article/what-is-asil-d>.
- NASA. 2020. *Program Project Management Development Process (PMDP) Handbook*. <https://www.scribd.com/document/364031766/NASA-Program-Project-Management-Development-Process-Pmdp-Handbook>.
- NIST. 2023. *AI RMF PLAYBOOK*. <https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf>.
- Palantir. 2023. *Palantir AIP for Defence*. <https://www.palantir.com/aip/defense/>.
- Rešlová, P., and M. Sassòli. 2022. *Meaningful Human Control in Autonomous Weapons*, Charles university.
- Roberts, P. 2019. "The Future Conflict Operating Environment Out to 2030." *RUSI Occasional Paper*. www.rusi.org.
- Royal Navy. 2016. *BR3(1) The Divisional and Regimental System*. 4th ed. Vol. 21. Ministry of Defence.

- Royal Navy. 2019. *BR3(1) Chapter 20 Discipline Policy and Process*, Ministry of Defence.
- Safety of Autonomous Systems Working Group. 2022. *Safety Assurance Objectives for Autonomous Systems*. <https://www.tpgroupglobal.com/>.
- Sassóli, M. 2014. "Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to Be Clarified." *International Law Studies* 308:308. <http://www.hrw.org/reports/2012/11/19/losing-humanity>.
- Sharkey, N. 2016. "Staying in the Loop: Human Supervisory Control of Weapons." In *Autonomous Weapons Systems*, 23–38. Cambridge University Press. <https://doi.org/10.1017/CBO9781316597873.002>.
- Sparrow, R. 2015. "Twenty Seconds to Comply: Autonomous Weapon Systems and the Recognition of Surrender." *International Law Studies* 91:725. <https://digital-commons.usnwc.edu/ils/vol91/iss1/20/>.
- Swedish Government. 2020. *Statement by Sweden at the CCW GGE on Lethal Autonomous Weapons (LAWS)*, Geneva. 21–25. September 2020. https://documents.unoda.org/wp-content/uploads/2020/09/200921-Anforande-LAWS-human-control_.pdf.
- TESLA. 2021. *Autopilot and Full Self-Driving Capability*. <https://www.tesla.com/support/autopilot>.
- Thurnher, J. S., and Schmitt, & Thurnher. 2013. "Out of the Loop, Autonomous Weapon Systems and the Law of Armed Conflict." *Harvard National Security Journal* 4 (2): 256. <https://doi.org/10.2139/ssrn.2271158>.
- UK MOD. 2022. "Ambitious, Safe, Responsible Our Approach to the Delivery of AI-Enabled Capability in Defence."



DSPACE

<https://dspace.org/>

Operating itself safely: merging the concepts of 'safe to operate' and 'operate safely' for lethal autonomous weapons systems containing artificial intelligence

Spayne, Peter; Lacey, Laura; Cahillane, Marie; Saddington, Alistair J.

2025

Taylor & Francis

<https://doi.org/10.1080/14702436.2024.2415712>

Spayne P, Lacey L, Cahillane M, Saddington A. (2025) Operating itself safely: merging the concepts of 'safe to operate' and 'operate safely' for lethal autonomous weapons systems containing artificial intelligence. Defence Studies, Volume 25, Issue 1, 2025, pp. 98-132