J. RESPONSIBLE AI FOR DEFENCE (RAID) TOOLKIT

MEASUREABLE ELEMENTS METHODOLOGY PAPER

BACKGROUND INFORMATION ON THE DRAFTING OF THE RAID TOOLKIT MEASUREABLE ELEMENTS

Contents

Part One. Introduction.

Many States, including Australia, have not yet released a formal framework of values and principles to address the design, development, or use of military AI. This should not however be viewed as such States not having a position, even potentially the substrate of a coherent informal framework to manage AI. For instance, in the case of Australia they have already publicly acknowledged a sufficient diverse range of elements to underpin a formal framework – if they were brought together in a formal structure.

This paper seeks to outline how the measurable elements in the RAID Toolkit were selected, and how they seek to ensure that the harmonised set of values and principles comprises those that can be measured. That is, a software or hardware designer could provide a metric or measurable output to put each element into practice during the design and development of their AI technology. In addition to enabling the incorporation of these values into design, such measurable elements will also enable the recording and future certification of systems for eventual adoption and use by militaries, as well as applying to the system as a whole within the proposed use context of the technology. In this way, this proposed list also aligns with the ISO/IEC/IEEE 24748-7000 Standard, which articulate a context-specific analysis of the values that underpin the AI system use.

Further, given the regulatory framework for responsible military AI is being built as the uncrewed airplane is in the air, so to speak, this paper's proposed rubric of elements also provides a methodology to continue to build upon the proposed core and harmonised elements in a way that can be adjusted by governments and industry in their application to AI technologies. Thus, as governmental requirements of industry change compliance with core legal and ethical frameworks can be demonstrated utilising this harmonisation methodology. This can be achieved by assessing whether the existing rubric contains the underpinning values or characteristics required of the particular system/s, and whether or not additional elements are required to address the particular use case or need.

A separate, but similarly vexed problem relates to the interoperability of AI military systems between likeminded States, where similar, but not identical value and principles frameworks are used to outline whether a particular AI system is capable of responsible use. By analysing prominent frameworks, and identifying the similarities in the underpinning values, principles or characteristics between them, a consolidated list of elements can ensure a level of interoperability between systems, when a developer is approaching their design of these systems. The outcome will then be an ability to demonstrate to a State, depending on its AI framework, how their system addresses their values requirements – and to what standard.

With this in mind, and also noting the rapid rate of development of and change to military AI frameworks, the proposed list of harmonised elements is far from complete. This list of elements has been derived from ten select military AI frameworks.[1] These frameworks have been selected because the proliferation of AI frameworks designed to address military AI use, as well as general adoption of AI technologies, mean that it is practically (and feasibly) not possible in this paper to address and incorporate all of them. Because the use is for Australian industry, they have been selected having regard to extant Australian principles and Australian key allies as the focus of the harmonisation; noting that such an activity can be undertaken with a focus upon any State.[2]

It is equally important to note that there is a difference in the purpose, approach and value proposition of frameworks seeking to regulate civilian use of AI systems, with some of the military uses of AI. For instance, the concept of Azimov's 'do not harm' law of robotics may align with a military 's requirement to comply with domestic

[1] See Part Two for a comprehensive list of which frameworks were selected.

[2] The authors wish to reiterate that this list does not represent the views of the Australian government. Key allies selected include: United States and the United Kingdom; NATO, based upon Australia's status as an Enhanced Opportunity Partner, including cooperation in NATO operations, such as the ISAF alliance in Afghanistan (NATO, *Relations With Australia*, https://www.nato.int/cps/en/natohq/topics_48899.htm). New Zealand does not, to date, have a publicly available military AI strategy: Moses 2021.

workplace safety requirements, but it simply does not align with a military's requirement to use force to respond to a threat (Sorrell, 2017). Equally, the *lex specialis* of the laws of armed conflict means that there is a different legal (and ethical) framework that will dictate how the system will operate in its specific context. A comprehensive framework for use by the military must therefore be capable of handling heterogeneity in AI (such as technical specifications, environment, and complexity) and their intended use.

This first iteration of the harmonisation of frameworks is designed to demonstrate the suitability of applying a singular list of harmonised elements in the design and development of AI technologies, and would be expected to evolve and adapt, depending on the potential future use of the list.[3] While seemingly an exercise in semantics, the reframing of the concepts into a single rubric is valuable in identifying a method for enhancing interoperability in coalition operations, and allowing AI developers to adhere to a framework that makes their product more universally acceptable and therefore potentially more marketable.

Part Two. The Methodology

The definition of AI

To begin with, it is important to note that there are many competing definitions of what is meant by the term AI. It could be for instance:

- *a collection of technologies able to solve problems and perform tasks without explicit human guidance (CSIRO, 2020).*
- *a thing used to perform tasks and solve problems that, if done by humans, would require thinking.*
- *a central technology that enables cognitive-like functions in a robotic and autonomous system (RAS) (Commonwealth of Australia, 2020).*

[3] A further use case for such an approach could be in the use of AI technologies within coalition environments, enabling cooperating States to rely upon coalition force AI technologies, within certain defined limits.

It is sufficient for the purposes of this rubric to rely upon the existing definitions of AI found in the Australian AI Action Plan, noting its broad alignment with the OECD Council on Artificial Intelligence definition (albeit any of the general definitions or explanations of AI would be sufficient for the purposes of this paper).[4]

There is currently no single, publicly available definition of what military-AI might be in the Australian context.[5] Importantly, for the purposes of this paper, while there are a breadth of military systems that will rely upon computers to make decisions that were previously the preserve of humans, there is no reason to consider what AI is in the military domain to be any different to AI the civilian domain. What is important, is that the use of certain AI in the context of the military domain will trigger ethical and legal obligations, and thus require consideration by a human utilising an appropriate responsible AI framework, prior to their use.

---

[4] While not a framework eliciting principles, the *Australian AI Action plan* provides an Australian government definition of what constitutes AI, (Commonwealth of Australia, 2021; OECD, 2019.)

[5] Australia does not have a single definition of AI included in its publicly available frameworks, but in addition to the definition in the *AI Action Plan* definitions include:

- Commonwealth Scientific and Industrial Research Organisation (CSIRO): 'Artificial intelligence (AI) may be defined as a collection of interrelated technologies used to solve problems autonomously and perform tasks to achieve defined objectives, in some cases without explicit guidance from a human being. Subfields of AI include machine learning, computer vision, human language technologies, robotics, knowledge representation and other scientific fields. The power of AI comes from a convergence of technologies.' (Hajkowicz SA , Karimi S, Wark T, Chen C, Evans M, Rens N, Dawson D, Charlton A., Brennan T., Moffatt C., Srikumar S., Tong K, 2019; CSIRO, 2019)*;*
- ADF: *'*AI is a collection of interrelated technologies used to solve problems and perform tasks that, when humans do them, requires thinking.' (Commonwealth of Australia, 2020; cited in Scharre, 2017; and Royal Australian Navy, 2021)*;*
- Australian Army: offers a fused definition, 'A collection of techniques and technologies that demonstrate behaviour and automate functions that are typically associated with, or exceed the capacity of, human intelligence,' (Commonwealth of Australia, August 2022).

Where NATO or OECD have been more specific in their regulatory approach, looking at limited class of AI systems, this rubric proposes that such classes of capability are merely one small part of AI that must be considered in terms of military use. Furthering the Klonowska (2020) model of co-production of hostilities (related to the obligation to conduct a legal review of that technology), this paper considers that any capability that has AI that supports, implements, or otherwise subsumes part of the human decision-making processes should be actively assessed for compliance with the responsible AI framework prior to its use (Klonowska, 2020).

This approach is intended to address the consideration of the military use of AI capabilities. It is not limited in its approach to only the use of AWS, nor is it intended to exclude use in garrison support or other domestic military contexts. That being said, there are fundamental shifts in the considerations, within the rubric, for the use of AI from both legal and ethical perspectives during certain military uses, such as armed conflict. For instance, we assume that international humanitarian law applies to capabilities intended for use in military operations;[6] as well as other legal obligations depending on the specified use case.[7]

Types of regulatory methods

While there are multiple approaches that can be taken when creating a set of measures designed to record the efficacy or effectiveness of a system, the elements described in this paper do no seek to favour one methodology above another. Rather, these elements seek to identify what the underpinning issues requiring measurement are; and leave the specific methodology to demonstrate standards compliance for the specific use case. There will be multiple methods to measure each element, depending on the design methodology adopted by the designers, or

[6] To the extent applicable to the relevant classification of conflict (for example, whether the laws of occupation, or the laws applicable to conflict or an international or non-international character are relevant).

[7] For example, the laws guiding peacekeeping operations will differ from those dictating how the capability must perform in a maritime counter-terrorism operation.

articulated by the State, and – most relevantly – the articulated used case for the capability, but the base element will remain the same.

A recent comparison of cybersecurity regulation to the methods of safety regulation in high-hazard industries reveals the same underlying approach for how AI technologies for use in the military must be regulated (Dempsey, 2022): a combination of regulatory methodologies will be necessary to address specific, context-specific AI use. The three primary regulatory methods that can be employed include: performance-based regulation, which require a specific, measurable output in performance from the capability; prescriptive regulation, which mandates a particular solution such as specifying a type of style of technology which may be used in a particular situation; and management-based regulation, which direct particular processes must be followed by regulatory entities (National Academies of Science, 2021). Accordingly, the articulation of measurable elements will enable one of the above regulatory methods to be applied, depending on the context for use.

Finally, the drafting approach applied here – specifying measurable elements but not directing regulatory methodology – provides longevity and flexibility to the adopted approach. It does not specify the actual method required to be undertaken; but can specify what needs to be measured as a result of the activity. In this way, the elements build flexibility to adjust depending on risk, context, environment, and legal framework applicable for use of the capability. It also provides flexibility to be updated in terms of step-changes in technology, unforeseen emerging and disruptive technology threats, and the general speed of relevance.[8] This flexibility can be applied while the constituent principles and values that form the elements, which will not change as rapidly, can endure.

Critically, this paper does not attempt to prescribe how the measures are to be applied. The choice of which measure to use, how they are to be applied

[8] For an analogous discussion relating to legal regulation and new technologies, see Easterbrook (1996).

(individually, or in combination), and the priority and weighting attached to each measure are all capable of manipulation to suit the specific requirements of the user. As such, this methodology allows for subjectivity in terms of context when assessing a capability for compliance using States' requirements. This methodology allows States to tell AI designers or developers what they need to achieve, for each measure, or combination thereof, to suit the States' needs.

A recent survey of the adoption of human-based values into software design identified 51 different processes for operationalising human values in software (Shahin, Hussain, Nurwidyantoro, Perera, Shams, Grundy & Whittle, 2022). They describe each of these 51 processes articulate a different methodology for identifying human values and translating them to accessible and concrete concepts so that they can be 'implemented, validated, verified, and measured in software'. The implementation of responsible AI also requires regulatory methods applied to the hardware; and risk-mitigation approaches to the context in which they are to be used. Multiplied across the spectrum of military operations, it is apparent that there is a potentially indeterminate number of differing processes that will be applicable. The regulatory method should therefore be capable of application in a case-specific manner, but by reference to standard values. This is achieved by translating those values into elements capable of bespoke implementation, validation, verification and measurement.

The methodology for deriving the elements.

The methodology builds upon the analysis contained in the Stanley-Lockman's 2021 *Responsible and Ethical Military AI* CSET Issues Brief, and Copeland and Devitt's 2022 assessment of Australia's approach to AI governance in Defence. This rubric, however, seeks to not only identify convergences and divergences in framework principles, but also proposes that the elements put forward by States need to include measurable elements that can be actioned by capability creators. Measuring and mitigating AI risk is not limited to the software being legally and ethically compliant, but also the accompanying hardware, analysed within the use-case context (Ezeani et al, 2021).

Accordingly, the elements are considered to address the underpinning characteristics and requirements of both the legal frameworks relevant to the particular use case, and include those key ethical theories flagged in the frameworks harmonised during this exercise. For instance, In the context of the military use of AI weapons, means and methods of warfare it has been assumed that the ethical framework/s adopted for use during armed conflict have been subsumed by states into specific legal obligations.[9] Just war theory, utilitarianism and consequentialism, while all important individual ethical frameworks treated separately in some standards processes (like, for example the ISO standard), have been incorporated into extant legal obligations for the use of these capabilities; or articulated in the higher-order ethical AI frameworks.

The actions of military personnel attract State responsibility; and are required to comply with numerous international and domestic legal and ethical obligations. These systems have therefore been designed to provide assurance that compliance can be achieved by the human actors within the military machine, even in the most difficult of human endeavours: the controlled and systemised use of violence to achieve the ends of the State. Conversely, there are no 'systematic software-engineering methods that detail how to define, refine, and monitor human values throughout the software-development lifecycle' ([1]Whittle, Ferrario, Simm & Hussain,

[9] Law and ethics are related and supporting concepts. In most militaries, acting lawfully is the minimum standard of ethical behaviour. In the Australian context, every Australian Defence Force (ADF) member is obligated to comply with Australian and international law. The ADF defines ethics as 'moral principles or standards of acceptable behaviour by which any particular person is guided.' It compels them to ask, 'Is it the right thing to do?'. Laws give practical effect to ethical principles and create mechanisms for enforceability and accountability. Ethics also guide the understanding of law, particularly in situations of uncertainty where lawful actions may not always be ethical. In a military context, and particularly during combat, both law and ethics recognise the centrality of humans as responsible and accountable actors. Ethical principles are the foundations upon which many laws are based. The legal concepts of *jus ad bellum* – the legal basis for States going to war – and *jus in bello* – the legal basis for conduct in war – give effect to the just war ethical framework. Similarly, International Humanitarian Law, or the Laws of Armed Conflict (LOAC), are a balance of the principles of military necessity and humanity resulting in significant protections for civilians, hors de combat and to a degree, combatants, through the rules of distinction, proportionality, and precautions in attack.

2020). Accordingly, the provision of a systemised rubric of values that can be readily adapted into the software (and hardware) lifecycle would be beneficial to be adopted by militaries, as well as to guide AI designers or developers, to ensure the appropriate translation of technology – at its highest – in situations of armed conflict. Put simply, values-based approaches still require measurement.

In the same way that there has been work undertaken to identify the values that should underpin use of autonomous systems used in armed conflict, there has been work undertaken to develop international standards relevant to AI capabilities. These standards, specifically the IEEE P7000 standard (and the ISO developing standards which adopts the IEEE methodology), incorporate values and principles into them as a stage of the design process, but do not provide assistance in delimiting or translating these values or principles into elements that can be readily measured or demonstrated in the design process. These standards create a process to translate the values and standards into systems and software engineering, applying engineering design practices like value-sensitive design.

This proposed rubric does not purport to replace such methodologies, nor does it purport to be a revolutionary approach to engineering design. Rather it proposes that in the design and development of military capabilities, the set of values that require measurement to ensure the capability can be deployed responsibly, should not change dramatically from capability to capability. Thus, a standing rubric of measurable elements, building upon the higher-order values frameworks and principles already in existence, should be capable of use by capability builders and users to ensure consistent compliance with ethical and legal standards by military AI.

Other considerations relevant to note for the below elements harmonisation and derivation activity are:

- *The term 'principle(s)' is used in its ordinary sense, using the Oxford Dictionary's definition of 'a fundamental truth or proposition that serves as the foundation for a system of belief or behaviour or for a chain of reasoning'; and which has been identified as part of a values, ethics or legal framework, required for AI to be responsibly fielded by a military.*

- *The term 'element(s)' is used to describe a principle required for AI to be responsibly fielded by a military, which is capable of implementation, validation, verification and measurement.*

'Responsible AI' as the headline principle

There are numerous different ways to conceptualise what the driving principle for a scheme that can operationalise legal and ethical principles for AI might be. We have selected the term 'responsible' AI, not simply because it is reflective of the current *zeitgeist* in AI frameworks, but because it includes in its essence the requirement to be lawful and ethical, and deployed in a way that complies with appropriate level of human control (US Department of Defense, 2022; Richards, Boulanin, Goussac & Bruun, 2020; Stanley-Lockman, 2021). Responsibility – such as the various modes of criminal responsibility, or as articulated in various ethical epistemologies – can take many different forms, but is intended to reflect a level of compliance commensurate with the assessed risk.

This also reflects the shift in conversation in the international domain to a focus on behavioural limitations (i.e. responsible behaviours or responsible AI), whether in conjunction with or in replacement of binding arms control regulation, as seen in the *Convention on Certain Conventional Weapons – Group of Governmental Experts on Lethal Autonomous Weapons Systems* ('CCW GGE on LAWS') 2022 discussions; as well as in recently released State military AI frameworks, and indeed, in this volume on responsible AI in the military (Atcheson and Lytlak, 2022).

The concepts of assurance, governance or regulation all play a part in how this scheme may be used and incorporates other considerations such a human-machine interaction and safety frameworks. However, fundamentally, when these parts are added together, this system is intended to demonstrate that the AI, when used by the military in a specific context, is capable of legal and ethical compliance.

Part Three. Elements for Responsible AI

The below rubric contains the 12 selected elements, based upon an assessment of AI frameworks considered relevant to the design, development and use of military AI capabilities (with a focus on the Australian context). Elements were reviewed to assess coverage, whether individually, or in combination, to comprehensively address the values and principles relevant to military use of AI. As discussed above, it is a rubric of elements that the Australian military can use to regulate their study, adoption, acquisition or use of AI, and Australian defence industry can use to guide their responsible design of military AI. It can later translate to certification by the capability users, should they adopt such elements. In any event, the elements represent a rubric of measurable principles that can be recorded and communicated to procurement agencies, and readily mapped to an organisation's certification processes for a specific AI use case. They extend in applicability across the capability life cycle, from the design and development phases to the introduction into service, deployment and disposal phases and thus will require application to the components of an AI system, including the system's intended objects, targets and operating environment.

Responsible

Human beings should exercise appropriate levels of judgment and care; and remain responsible for the development, deployment, use, and outcomes and consequences of AI systems.

This is the headline principle for this rubric – just as it is the cornerstone of many AI frameworks – and addresses certain key concepts not addressed by other elements. Specifically, this element sets the requirement for the system design, use and consequences to be connected to a responsible human; and that the approach to the design of the system has the ability to be responsible insofar as it is exercising judgment in the manner in which the system will be designed and operated. This element is inextricably linked with the measure of accountability, albeit it effects and is affected by all the other elements. This element features in most AI frameworks.

While Australia does not have a publicly released framework, it has supported the importance of this concept in multiple fora.[10]

The key internal practice available to States to ensure responsibility is part of an AI capability is 'baking' human intent into the code. This practice is measurable in a number of ways, ranging from code review (premised on design specifications) to statistical analysis of Testing, Evaluation, Verification and Validation (TEVV) applied to code function and capability effects. Other examples of methods to measure this element include: training, education, orders regarding activation; use of blockchain to record which commander/use is temporally responsible for the AI capability at any given time, assessment, testing and certification processes applied to operators and commanders prior to being permitted to operate or command operation of an AI. Clear articulation of responsibility for the AI and attendant risks at each stage of the AI capability life cycle – including where appropriate apportioning 'levels of responsibility' across its design and use phases, to individuals including operational commanders (Devitt, K., Gan, M., Scholz, J. & Bolia, R., 2021).

Human Factors and Ergonomics (HFE) methods relevant to this element include: 'task analysis, cognitive task analysis, process charting, situation awareness assessment, trust assessment, mental workload assessment, teamwork assessment, interface analysis, usability evaluation, design, systems analysis, risk assessment, and accident analysis' (Salmon, P., King, B., McLean, S. & Read, G., 2022).

[10] The list includes multiple frameworks, such as: NATO, 2022; UK 2022; Devitt, K., Gan, M., Scholz,J. & Bolia, R., 2021; Australian Defence Force, 2022; Australia Group, March 2022; United Kingdom Ministry of Defence, June 2022; US Department of Defense, 2021; Finland, France, Germany, the Netherlands, Norway, Spain, and Sweden, 2022; Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States, 07 March 2022.

Accountable

People responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and human oversight of AI systems should be enabled. [11]

The use of the term 'people' also recognises the need for legal personality: in certain cases, a 'person' for legal purposes could include a corporation with responsibility for the development of the capability. This element is inextricably linked to the element of responsibility.

The primary measures of accountability will be State prescriptions and military processes. State prescriptions (including legal frameworks, orders, doctrine, and policies) detail where accountability for using a capability rests. Military processes (often a component or result of State prescriptions) provide accountability through prescription or consequence of practice. For instance, an example methodology to measure accountability for the use of AI, may be formulated in and recorded by a blockchain enabled handover/takeover of command and control.

Understandable

AI-enabled systems, and their actions, decisions, behaviours, intention and predicated impact, must be appropriately understood by relevant individuals, with mechanisms to enable this understanding made an explicit part of system design. [12]

There are two components to this: the technical characteristics of the AI, and the capability and knowledge of the human, which implies an ability of the operator to

[11] Found referenced in: Argentina, Ecuador, Costa Rica, Nigeria, Panama, the Philippines, Sierra Leone and Uruguay, July 2022; United Kingdom, July 2022; Commonwealth of Australia, 2020.

[12] This element is found in the UK, Finland and Australian models.

'have appropriate awareness of the AI capabilities'.[13] This element is strongly tied to the elements of explainability and reviewability.

This element can be measured by testing of technical knowledge of AI operators and commanders on the AI, physical and virtual testing and certification based upon anticipated use and use limitations, test transparency of the AI capability from activation to operation to post activity review, HFE methods and assessment of the accuracy of user mental models of the AI capability.

While the goal of explainability is understanding (Henin & Le Métayer, 2021), these two elements are different in that their temporal positioning. Understanding is required prior to use whilst explainability only becomes relevant once the AI has been used. A further difference is that understanding requires a knowledge of how the AI works, including the range of outcomes and capability parameters from the prior information known about the AI, whilst explainabilty focusses on a specific outcome of the operation of the AI, and being able to explain that outcome.

Explainable

For any given use of the AI, the operation, the output, and the impact are intelligible.[14]

Design measures to measure this element are those that ensure the AI is 'intrinsically interpretable'. For example, undertaking TEVV testing to assess design measures. Other methods would include applying purpose-built AI to 'interpret' the AI system being validated; as well as standard HFE methodologies.

[13] Devitt, K., Gan, M., Scholz, J. & Bolia, R., 2021

[14] Explainable is found in the NATO, Finland, and Australian models.

Reviewable

When AI functionality is exercised the relevant aspects of the exercise of that functionality must be discoverable, recordable; and then capable of being audited to determine if AI functionality operated as intended when it was engaged.[15]

Discoverability, recordability and auditability of the algorithmic processes in the exercise of the algorithmic functionality will require integration with a military's extant acquisition and design processes; and have regard to the standards of information and evidence required for potential future investigations and interrogation in the event of misuse of accident. This element is strongly linked to understandability and explainability.

Reliable

AI systems and their algorithms should reliably operate consistently in accordance with their intended purpose and explicit, defined use cases. Reliability is the extent to which a system does or does not fail.[16]

This element will often be considered along with predictability (which measures the output from the operation of the AI). While some frameworks include the requirement for robust testing and assurance, across their life cycle to demonstrate safety, security and robustness (see NATO, 2022 and US DoD, 2022), we consider this is an inherent requirement to meet this element and thus is not required in its definition.

This element deals with biases; including considering biases in the data set itself, and the structures and systems in which they are being modelled. It can be

[15] This element is found in the Australian and UK models.

[16] This element is found in the NATO, US, Chinese, Finnish and Australian models.

measured using physical and virtual testing of to identify statistical failure rates; and through TEVV that addresses inputs and context for its specific use case. It can also be measured using adversarial testing and assessment against standards of performance. HFE methods to measure this element include, 'cognitive task analysis, process charting, human error identification, situation awareness assessment, trust assessment, mental workload assessment, teamwork assessment, interface analysis, usability evaluation, performance time prediction, design, systems analysis, risk assessment, and accident analysis' (**Salmon, P., King, B., McLean, S., & Read**, G., 2022).

Predictable

The actions and effects of the operation of the AI capability should be anticipated and intended. [17]

This element is often considered alongside reliability (which measures the ability of the system to undertake its set functionality), but will be important for other elements such as understandability and controllability. Predictability can be broken into three parts:

- *the degree to which system's technical performance is or is not consistent with past performance;*
- *the degree to which any AI or autonomous system's specific actions can (and cannot) be anticipated; and*
- *the degree to which the effects of employing an AI system can be anticipated (Arthur, 2020).*

Measuring this element could be achieved through TEVV, testing of data sets for hygiene (and in particular for bias mitigation), the conduct of adversarial testing, virtual/simulation and live field training. The focus of any such measures will be on the effects from such activities and will have a strong statistical component. The

[17] Austria, Finland and Australia include this element.

concept of predictability is often referred to as the 'black box' problem with AI; and thus the development and application of 'standardized metrics to grade predictability and understandability' to test against will likely to be a requirement to appropriately measure this element.

Controllable

The AI must operate within a responsible chain of human command and control and be capable of human influence as designed.[18]

The concept of controllability refers to the requirement for the AI to be capable of being controlled in accordance with human intent. It does not necessarily mean that the AI requires a human in or on the loop in relation to its operation; rather it means that within the considered design specifications, the ability for human involvement with the capability aligns to the design specifications within risk tolerances (such as the operation of kill switches for certain use cases, or continuous tethers for others).

Controllability is linked to the capability of the AI to be used in compliance with any extant regulatory frameworks and will affect the ability to satisfy other elements such as understandability or safety (Dempsey, 2022). Synonymous concepts include the AI being 'manageable' or subject to 'appropriate human judgement' or 'human control'' (despite the extend and limits of this principle being ill-defined and not consistency understood. In the Australian context, this element refers to the systems that are placed over the system. Measuring this element requires demonstration of compliance with the control mechanism, whether that be physical, software or hardware related, or derived through a system of control for the AI capability when deployed within the specific use-case environment. It subsumes the concept of 'governability', which is variously described as the requirement for AI systems to function as intended, and to be responsive to deactivation when the systems demonstrate unintended behaviour.

[18] This element has been articulated in the Austrian, Finnish, Australian and GGE States models.

HFE methodologies that could be utilises include: physical HFE methods, task analysis, cognitive task analysis, process charting, human error identification, situation awareness assessment, trust assessment, teamwork assessment, interface analysis, usability evaluation, design, systems analysis, risk assessment, and accident analysis (**Salmon, P., King, B., McLean, S., & Read**, G., 2022).

Compliant

> *The activation and operation of the AI capability must be compliant with the applicable law, ethics, the governance framework, and the broader system of control.*[19]

This element is potentially reliant on a number of other elements depending upon the type of AI; for instance, controllability, understandability and predictability are highly relevant to AI in means and methods of warfare. The element incorporates the requirement for legal compliance having regard to the particular use case of the capability and system within which it integrates, hence why the use of 'law' as a separate element would be redundant.

Methods to measure this element include:

- *Thorough TEVV, with a focus on identifying and assessing effects for predictability and controllability;*
- *Assessing the extent to which appropriate values and standards have been designed into the AI capability;*
- *Assessing the extent to which inputs – such as data – have been appropriately validated. For instance, ensuring good data hygiene can reduce output bias; algorithm TEVV can avoid algorithmic bias; and value sensitive design can offset automation bias.*

[19] Argentina, Russia, Austria, UK, Finland, Australia, and the GGE States refer to this principle in their frameworks.

- *Use of compliance toolkits, such as the AI Fairness 360 toolkit, to identify biases and discrimination (Devitt, K., Gan, M., Scholz, J. & Bolia, R., 2021).*

HFE methodologies to measure this element could include: human error identification, situation awareness assessment, trust assessment, mental workload assessment, design, systems analysis, risk assessment, and accident analysis (Salmon, P., King, B., McLean, S. & Read, G., 2022).

Integrated

> *The ability of the AI to integrate into the AI capability, any system it controls or supports, into the HMI, and into the broader system of control.* [20]

This is perhaps the least expected element in the rubric. It has been included to emphasise that not only must an AI be bounded (controllable), but that it must be capable of working within a system, whether as decision support and working to or for a human operator, or as an integrated component on a system (such as a weapon or means of warfare), and within the broader system of control. Unsurprisingly, this element should be considered alongside controllability, however it will be important for human interaction with the AI (understandability, explainability, reviewability), as well as statistical analysis of the operation of the various system layers (reliability, predictability) relating to the AI.

The integration of the AI into the various systems will come from measures derived to assess other elements but would benefit from TEVV to identify effects of AI functionality when operated with the various systems it is expected to interact with. As such the continuation of such testing once the AI has been accepted into service through virtual or simulated training, as well as live training in various levels of controlled to semi-uncontrolled environments would be necessary to measure this element.

[20] This element is found in the Australian Defence Science and Technology Group report.

Safe

The AI capability should not pose unreasonable safety risks, and should adopt safety measures that are proportionate to the magnitude of potential risks. [21]

While the measurement of all these principles relates to a risk management activity, the management of safety risk is a specific and different element requiring inclusion. Safety is a fundamental element (alongside controllability, predictability and understandability) to create trust in the users of AI. As AI advances it will likely become enmeshed in the security element as a reflection of safety from intended uses as well as safety from malicious intrusions.

There exist a wide variety of safety risk measurement and management methodologies, but the safety of the AI system can generally be measured using recursive testing, evaluation, verification and validation process, where systems (both active and non-active learning). Continual and iterative testing and certification will support the measurement of the system's safety (Arthur, 2020).

Secure

The AI capability, including its data, must be physically, cyber and electromagnetically secure. [22]

It addresses privacy protection as well as security more generally. Security will likely be viewed in tandem with safety with respect to malicious intrusions and will affect user and societal mistrust if it cannot be maintained. It can be addressed through physical, cyber and electromagnetic measures.

---

[21] The US, UK and Finland include this element in their frameworks.

[22] A version of this element is found in the Chinese, Australian, Argentinian, Russian, Finnish, US and GGE frameworks.

Part Four. Principles from frameworks not included

The above rubric represents those elements considered to address the central legal and ethical issues applicable to the design and development of AI systems intended to be applied to specified military use cases. There remain, however, principles articulated in the assessed AI frameworks that have been excluded from our proposed scheme. These excluded principles, and the reasoning for their exclusion are detailed below. There are two broad categories of excluded principles: the first being those that could have been included but were not because their consistent characteristics are addressed by the accepted elements; the second category has been excluded on the basis that they represent a synonym for an accepted element included in the list.

Lawfulness

Somewhat controversially for a scheme intending to incorporate the legal and ethical design considerations for the use of AI systems, an element labelled 'law' has not been incorporated as a standalone element. Some frameworks or policies specify law as a standalone principle.[23] Some frameworks, such as the US and UK, do not list 'law' as a standalone principle, rather, it is a separate, and fundamental consideration that does not require incorporation as a principle, because compliance with the law is considered a foundational and non-negotiable requirement in the use of an AI system. Such consideration is separate to the incorporation of values or ethical principles, which can be adopted in varying degree without breaching a State's fundamental legal obligations.

Lawfulness has not been incorporated as a standalone element precisely because it is central to the endeavour being undertaken by enunciating the framework in this first instance: the list of elements is designed to provide a method to demonstrate how the law is being complied with and incorporated into every element of design.

[23] Of Australia, NATO and China, for example.

On that basis, it does not require its own element, but rather is subsumed into the broader element of compliance which covers compliance with the law as well as other requirements for compliance (such as ethical, system control, and policy).

Bias Mitigation

Bias mitigation is found in the NATO, UK, Argentinian joint proposal and Australian AI Ethical Principles. It is described as equitable in other frameworks; and essentially considers the need to implement strategies to mitigate the potential for harm caused by algorithmic bias caused by the AI system. The frameworks address concerns relating to the applications, its input data being protected from unexpected or unintended biased, representative of its operating environment, and its outputs achieving the desired effect. They also articulate a need for proactive bias mitigation processes to ameliorate these risks.

Bias mitigation, and the various risks it describes, are considered to be addressed principally by the element of predictability, as well as by the elements of controllability and compliance (insofar as compliance deals with the appropriate validation of input data). It is of course important to note, and perhaps why this element is not included in the rubric, that some biases are not necessarily inappropriate or require mitigation (for instance a bias to label children in situations of doubt from AI used to support targeting), but it would be important to ensure such biases are recognised and accounted for.

Human Centricity

The level of human control, and the nature of decisions that may be taken by a machine, rather than a human with appropriate agency, remains a controversial issue that has not been settled in many State's frameworks. This element is applied in the rubric primarily through the human participation focussed elements of responsibility, accountability, understandability and explainability. It will also feature strongly in controllability and compliance as these relate to human involvement and adhering to human responsibility.

Fairness

The Australian AI Ethics Action Plan articulates fairness as requiring that technologies should be 'inclusive and accessible; and should not involve or result in unfair discrimination against individuals, communities or groups' (Australia, 2020). While fairness is, in general, a noble pursuit in the use of AI, this civilian principle is not one that is considered always compatible with military use of AI. For instance, for those AI capabilities intended to be used in support of a means or method of warfare, international humanitarian law will be a relevant regulatory consideration. International humanitarian law allows for unfair decisions to be made, provided they are not arbitrary. An example of this, during situations of armed conflict, is that international humanitarian law permits persons to be interned based upon the risk profile they pose premised only upon their citizenship status. This is, objectively unfair but not arbitrary.

Accordingly, this principle is not one that is necessarily appropriate for use during the exigencies of armed conflict. That said, compliance with this principle to the extent it is possible, is desirable, and in some cases lawfully required. It will therefore impact upon certain military uses of AI. The principle of fairness underpins many international human rights law concepts; and should be applied to the extent it is not inconsistent with international humanitarian law principles.

This principle, to the extent it is relevant, is applied through the elements of responsibility and accountability.

Traceability

Traceability features in both the US DoD and NATO frameworks and are articulated as being appropriately 'understandable and transparent'. This principle is considered synonymous with the element of explainable, but complemented by the elements of reviewable and understandable. The methodologies proposed by these frameworks to achieve this principle, respectively, include 'review methodologies, sources and procedures' and 'transparent and auditable methodologies, data sources and design

procedure and documentation'. It is also a principle espoused by the Australia Group Proposal at the GGE on LAWS in 2022 (Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States, 2022).

These methodologies have primarily been subsumed into the rubric's elements of understandable explainable and reviewable.

Governability

The principle of governability is found in multiple military AI frameworks. This principle is considered to be addressed across the multiple the elements of controllability, understandability, explainability, predictability and reliability.

Trust

Trust is articulated in three of the Australian frameworks as a central principle for the use of AI systems. It is described in the Australia group GGE submission (2022) as 'the firm belief in the reliability, truth or ability of someone or something', and in the context of the Method for Ethical AI, it was described as a central facet, whereby, '[h]uman-AI systems in Defence need to be trusted by users and operators, by commanders and support staff and by the military, government and civilian population of a nation,' (Devitt, K., Gan, M., Scholz, J. & Bolia, R., 2021). Despite being a relatively intangible, personal, and difficult to measure concept, trust is central to the relationship between human and machine, and their interaction. This principle is considered to be the result of a responsible AI framework, and thus is not listed as its own element.

Responsibility is the corollary to trust insofar as a responsible system engenders trust in its users. It is a measure of success rather than a system requirement in and of itself. While trust measures require building in to the system design of military AI capabilities, in the same way they must be for all information technology systems, trust has not been included as an element on the basis that it is amenable to appropriate user interaction (understandability and explainability), a result of being a reliable and predictable system, that is capable of being controlled, is safe and secure, and satisfies user compliance requirements.

Manageable

Manageable is found in the frameworks of China (2020) and Australia (2020); and is considered a synonym of controllable. This element is considered synonymous with controllability and as such has not been included as a separate element.

Transparent

Transparency is linked to the concept of explainability. It features heavily in civilian AI frameworks, where a civilian population is entitled to certain rights in terms of privacy and protection from the actions of their government. However, this part of the principle is not necessarily relevant in terms of military AI use in armed conflict when considering the impact of AI against an enemy (albeit that it may be relevant in terms of use against the civilian population).

Transparency also suggests a need to be able to 'see' how an AI system reached a conclusion, however, we content that it is more relevant to ensure that the outcome of and in the case of convolutional neural networks it may not be possible to ever 'see' what the AI's process is, rather it is only possible to predict how the AI will perform. In this sense, the element of predictability is relevant.

Insofar as transparency relates to an ability to have 'awareness of an autonomous agent's actions, decisions, behaviours, and intention', then the element of predictability, reliability, understandability, explainability, and reviewability address this principle.

Risk Mitigation.

Both the Finnish and Australian frameworks suggest that 'risk mitigation' is a principle relevant to the use of AI. We contend that risk mitigation is more relevantly considered one methodology of regulation, that can be used to address the risks that manifest in the use of AI technologies across the relevant spectrum of elements and principles. As such this element would be addressed within the rubric by a combination of controllability and compliance, as well as safety and security.

Privacy

The Australian AI Action plan speaks to privacy in a number of terms. It describes the requirement to protect both privacy and security by upholding 'privacy rights and data protection', as well as in terms of 'the security of data' (Australia, 2020).

The elements of predictable, safe, controllable and secure address this principle.

Human, societal and environmental wellbeing

The Australian civilian principles consider that the AI systems 'should benefit individuals, society and the environment.' This principle is relevant to the military use of AI in peacetime and for traditional support roles such as administration and logistics, as well as peacetime activities more generally. This principle loses relevance when considered in the context of military operations. While there are legal limits on the impact of military operations on the environment, insofar as they may not be excessive, or cause wide-spread, long-term or severe damage to the environment, there is no lawful prohibition on causing environmental damage during the conduct of hostilities, provided it does not breach this criterion (*Additional Protocol I to the Geneva Conventions*, Articles 35(3) and 55(1)). It is also not unethical for a State, during the conduct of hostilities, to cause such damage. A similar difference in approach between military operations and civilian principles apply in terms of the other limbs of this principle.

Contestability

The concept of contestability by design, in terms of algorithmic decision-making outcomes is not a novel concept, however, is one that relates to the ability to interrogate and understand how an algorithm achieved a particular result. The concept of algorithmic contestability is grounded in the ability to articulate what a good or desired result is, and compare that to the outcome in the use case (Henin & Le Métayer, 2021). In the case of the Australian AI Action Plan, it describes the ability of an affected person (or community) to challenge an outcome of the AI system via a 'timely process' (Australia, 2020).

As this is part of the existing system in terms of military operations, this has not been added as an element. That is, the conduct of military operations attracts both individual, command and State responsibility for actions, as well as governmental and public scrutiny for activities that are undertaken. The relevant consideration in the design of AI military technologies is not that they can be contested, but that their design, development and use are capable of integration into extant accountability measures. A military must be capable of interrogating the actions of the capability as part of the system in which it was deployed, and this concept is addressed in the elements of explainability and reviewability.

Part Five. Concluding observations

It is relevant at this point to reiterate that the proposed rubric of elements is not intended to be a finalised list, nor is our attempt at creating this rubric considered to be more complete than the frameworks analysed as part of this review. It is also not a reflection of the Australian government position on these elements, but a proposed set of measurable elements that account for identified existing obligations and frameworks publicly released by the Australian government or its agencies/departments.

Instead, through the process outlined above, we have sought to create a framework representative of extant articulations of AI values and principles that can operationalise ethical and legal principles. The intent is to enable their incorporation into the design and development of AI systems, such that the success of the adoption of the underpinning values and principles of broader AI frameworks can be measured, recorded, and assessed.

This set of measurable elements can facilitate developers and designers in recording their actions utilising one or several methodologies which will enable them to effectively communicate how their capability can meet the values and standards of a particular State in an objective, rather than subjective manner. The level of compliance, and to what level the standards must be met, is subjective and will

necessarily reflect the level of risk a State is willing to accept in terms of a particular capability or context in which a system is designed to operate. This subjective level of risk acceptance – such as a 5% risk of target misidentification being acceptable, for example – can then be translated into an objective measure for developers to be required to meet or demonstrate.

Which underpinning principles become important for adoption will no doubt change as AI frameworks are further developed and refined; but this proposed methodology allows for the addition or deletion of the elements selected for inclusion in the scheme based upon such changes or developments. The critical concern relating to AI systems is that they are being built today, for use in the very near future, and need to incorporate still-to-be-decided operational frameworks. This scheme proposes a methodology that will allow legal and ethical issues to be incorporated as a design feature, rather than an afterthought; and for such systems to be capable of demonstrating compliance with the majority of AI frameworks currently in use.

References

Acheson, R and Pytlak, A. (2022, June 3) CCW Report, Vol 10, No 4, Autonomous weapons and questions of ethics, control and accountability', *Reaching Critical Will*. Retrieved 28 September, 2022, from https://www.reachingcriticalwill.org/disarmament-fora/ccw/2022/laws/ccwreport/16277-ccw-report-vol-10-no-4.

Agenda. World Economic Forum. (n.d.). Retrieved October 31, 2022, from https://www.weforum.org/agenda

Arthur, H., 2020. 'The Black Box, Unlocked: Predictability and Understandability in Military AI.' United Nations Institute for Disarmament Research Report, (22 December 2020). doi: 10.37559/SecTec/20/AI1

Australia, Canada, Japan, the Republic of Korea, the United Kingdom, and the United States*, Principles and Good Practices on Emerging Technologies in the area of Lethal Autonomous Weapons Systems* (7 March 2022), 1st Session of the UNGGE on LAWS.

Autonomous Weapon Systems in international humanitarian law - joint air power competence centre. Joint Air Power Competence Centre - NATO's Advocate to Air and Space Power. (2022, May 5). Retrieved October 31, 2022, from https://www.japcc.org/articles/autonomous-weapon-systems-in-international-humanitarian-law/

Commonwealth of Australia, (2020). *ADF Concept for Robotics and Autonomous Systems.*

Commonwealth of Australia (2020) (2). Department of Defence, *Australian Defence Force Concept for Robotics and Autonomous Systems.*

Commonwealth of Australia, Department of Industry, Innovation and Science, (2021). *Australia's Artificial Intelligence Action Plan*. Retrieved 10 September 2022 from https://www.industry.gov.au/dataand-publications/australias-artificial-intelligence-action-plan.

Commonwealth of Australia, (2022). *Australian Army RAS-AI Strategy*, V2.0. Retrieved 10 September 2022 from http://researchcentre.army.gov.au/sites/default/files/Robotic%20and%20Autonomous%20Systems%20Strategy%20V2.0.pdf.

CSIRO, *Artificial Intelligence.* Retrieved 15 September 2022, from https://www.csiro.au/en/research/technology-space/ai.

Dempsey, J. *Lawfare Blog*, 'Cybersecurity Regulation: It's Not Performance Based if It Can't Be Measured', (6 October 2022). Retrieved from, https://www.lawfareblog.com/cybersecurity-regulation-its-not-performance-based-if-outcomes-cant-be-measured.

Department of Defense: Responsible Artificial Intelligence Strategy and implementation pathway. GovWhitePapers. (n.d.). Retrieved October 31, 2022, from https://govwhitepapers.com/whitepapers/department-of-defense-responsible-artificial-intelligence-strategy-and-implementation-pathway

Devitt, K. and Copeland, D. 'Australia's approach to AI governance in security and Defence' in Eds M. Raska, Z. Stanley-Lockman and R. Bitzinger. (2022) AI Governance for

National Security and Defence: Assessing Military AI Strategic Perspectives, Routledge.

Devitt, K., Gan, M., Scholz, J. and Bolia, R., *A Method for Ethical AI in Defence, Department of Defence, Defence Science and Technology Group, Technical Report*, DSTG-TR-3786 (January 2021). Retrieved from https://www.industry.gov.au/data-and-publications/australias-artificial-intelligence-ethics-framework.

Easterbrook, F. "Cyberspace and the Law of the Horse," *University of Chicago Legal Forum* (1996) 207.

Enzeani, G.,et al, *A survey of artificial intelligence risk assessment methodologies*, Ernst & Young Trilateral Research Report (August 2021).

Field, M. (2021, May 20). *Was a flying killer robot used in Libya? quite possibly*. Bulletin of the Atomic Scientists. Retrieved October 31, 2022, from https://thebulletin.org/2021/05/was-a-flying-killer-robot-used-in-libya-quite-possibly/

Floridi, L, et al., "AI4People – An Ethical Framework for a Good AI Society", *Minds & Machines* (2018) 28.

Group of Government Experts (GGE) on emerging technologies in ... UN LAWS GGE. (2022). Retrieved October 31, 2022, from https://indico.un.org/event/37347/page/0

Government of the United States, *Responsible AI Guidelines,* Defence Innovation Unit, (2021). Retrieved 10 September, 2022, from https://www.diu.mil/responsible-ai-guidelines.

Hajkowicz S.A , Karimi S, Wark T, Chen C, Evans M, Rens N, Dawson D, Charlton A, Brennan T., Moffatt C, Srikumar S, Tong K, (2019*Artificial intelligence: Solving problems, growing the economy and improving our quality of life,* CSIRO Data61.

Henin C., & Métayer, D., (2021). 'A framework to contest and justify algorithmic decisions', *AI and Ethics*, 1:463–476

IBM, IBM's Principles of Trust and Transparency. Accessed 15 September 2022, from https://www.ibm.com/policy/trust-transparency-new/.

International Standards Organisation, ISO/IEC/IEEE 24748-7000 - *Systems and software engineering — Life cycle management — Part 7000: Standard model process for addressing ethical concerns during system design*. Retrieved 15 September 2022, from https://www.iso.org/standard/84893.html?browse=tc.

Klaudia Klonowska, K. (2022) 'Article 36: Review of AI Decision-Support Systems and Other Emerging Technologies of Warfare' (2020) *Yearbook of International Humanitarian Law* (23)123.

Letter dated 91/03/13 from the secretary-general addressed to the president of the Security Council, U.N. Doc. S/22393 (22 Mar. 1991). (2002). *International Terrorism: A Compilation of U.N. Documents (1972-2001) (2 Vols.)*, 507–507. https://doi.org/10.1163/9789004481398_054

Mohan, S. (2022, October 6). *Managing expectations: Explainable A.I. and its military implications*. ORF. Retrieved November 1, 2022, from https://www.orfonline.org/research/explainable-a-i-and-its-military-implications/

Joint Media Statement, Morrison, Scott and Payne, Marise, '*Australia to pursue nuclear-powered submarines through new trilateral enhanced security partnership'*, (16 Sep 2021). Retrieved from https://www.minister.defence.gov.au/statements/2021-09-16/joint-media-statement-australia-pursue-nuclear-powered-submarines-through-new-trilateral-enhanced-security-partnership.

Moses, J. et al, (2021, August 2021). 'New Zealand Could Take A Global Lead in Controlling the Development of Killer Robots – So Why Isn't It?', *The Conversation*. Retrieved 15 September 2022, from https://theconversation.com/new-zealand-could-take-a-global-lead-in-controlling-the-development-of-killer-robots-so-why-isnt-it-166168.

National Institute of Standards and Technology AI Risk Management Framework. U.S. Department of Commerce. (2022, August 18). Retrieved October 31, 2022, from https://www.commerce.gov/bureaus-and-offices/nist

NATO, *NATO Artificial Intelligence Strategy,* (22 October 2021). Retrieved 5 October 2022, from https://www.nato.int/cps/en/natohq/official_texts_187617.htm

OECD, *The OECD AI Principles*, Organisation for Economic Co-operation and Development (OECD), (2019). Retrieved 28 September 2022, from https://www.oecd.org/going-digital/ai/principles/

PwC, 'PwC's Responsible AI Toolkit in PwC' (2020). Retrieved 11 September 2021, from https://www.pwc.com/gx/en/issues/ data-and-analytics/artificial-intelligence/what-is-responsible-ai.html.

Richards, L., Boulanin, B., Goussac, N. & Bruun, L. SIPRI Report, *Responsible Military Use of Artificial Intelligence: Can the European Union Lead the Way in Developing Best Practice?,* (2020).

Rolls Royce, *The Aletheia Framework* (2020). Retrieved 30 September 2022, from https://www.rolls-royce.com/sustainability/ethics-and-compliance/the-aletheia-framework.aspx

Royal Australian Navy, Warfare Innovation Navy. *RAS-AI Strategy 2040* (2022).

Salmon, P., King, B., McLean, S., Read, G., *A framework of Human Factors methods for safe, ethical, and usable Artificial Intelligence in Defence* (July 2022), University of Sunshine Coast Report.

Scharre, P. *The Opportunity and Challenge of Autonomous Systems*, (2017) NATO.

Shahin, M, Hussain, W, Nurwidyantoro, A, Perera, H, Shams, R, Grundy, J & Whittle, J. (July, 2022). 'Operationalising Human Values in Software Engineering: A Survey', *forthcoming*, *IEEE Access Journal.*

Sorrell, T. (2017, March). 'Asimov's Laws of Robotics Aren't the Moral Guidelines They Appear to Be', 17 March 2017. *The Conversation*

Stanley-Lockman, Z. (2021). Responsible and ethical military AI. Taddeo, M. & Blanchard, A. (2022). 'A Comparative Analysis of the Definitions of Autonomous Weapons Systems', *Journal of Science and Engineering Ethics*.

Taddeo, M. & Blanchard, A. (2022). 'A Comparative Analysis of the Definitions of Autonomous Weapons Systems', *Journal of Science and Engineering Ethics*. https://doi.org/10.51593/20200091

UK Ministry of Defence. (2022, June 15). *Ambitious, safe, responsible: Our approach to the delivery of AI-enabled capability in Defence*. GOV.UK. Retrieved October 31, 2022, from https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence

UK Ministry of Defence. (2022, June 15). *Defence Artificial Intelligence Strategy*. GOV.UK. Retrieved October 31, 2022, from https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy

United Kingdom Proposal for a GGE Document on the Application of International Humanitarian Law to Emerging Technologies in the Area of Lethal Autonomous Weapon Systems (LAWS), (2022). Retrieved October 31, 2022, from https://documents.unoda.org/wp-content/uploads/2022/05/03032022-UK-Proposal-for-Mar-2022-LAWS-GGE.docx

United Nations. (2022, July 13). *Working paper submitted by Finland, France, Germany, the Netherlands, Norway, Spain, and Sweden to the 2022 Chair of the Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems (LAWS),* . United Nations. Retrieved October 31, 2022, from https://www.un.org/en/conferences/npt2020/documents

Whittle, J., Ferrario, M., Simm, W., and Hussain, W.,, "A Case for Human Values in Software Engineering," in *IEEE Software*, vol. 38, no. 1, pp. 106-113, Jan.-Feb. 2021, doi: 10.1109/MS.2019.2956701.

Working Paper - Draft submitted by Argentina, Ecuador, Costa Rica, Nigeria, Panama, the Philippines, Sierra Leone and Uruguay. Documents. (2022). Retrieved October 31, 2022, from https://reachingcriticalwill.org/disarmament-fora/ccw/2022/laws/documents