August 2021

Responsible and Ethical Military AI

Allies and Allied Perspectives

CSET Issue Brief

AUTHOR
Zoe Stanley-Lockman

Executive Summary

Since the U.S. Department of Defense adopted its five safe and ethical principles for AI in February 2020, the focus has shifted toward operationalizing them. Notably, implementation efforts led by the Joint Artificial Intelligence Center (JAIC) coalesce around "responsible AI" (RAI) as the framework for DOD, including for collaboration efforts with allies and partners.[1] With a DOD RAI Strategy and Implementation Pathway in the making, the first step to leading global RAI in the military domain is understanding how other countries address such issues themselves. This report examines how key U.S. allies perceive AI ethics for defense.

Defense collaboration in AI builds on the broader U.S. strategic consensus that allies and partners offer comparative advantages relative to China and Russia, which often act alone, and that securing AI leadership is critical to maintaining the U.S. strategic position and technological edge. Partnering with other democratic countries therefore has implications for successfully achieving these strategic goals. Yet the military aspects of responsible AI that go beyond debates on autonomous weapons systems are currently under-discussed.

Responsible and ethical military AI between allies is important because policy alignment can improve interoperability in doctrine, procedures, legal frameworks, and technical implementation measures. Agreeing not only on human centricity for militaries adopting technology, but also on the ways that accountability and ethical principles enter into the design, development, deployment, and diffusion of AI helps reinforce strategic democratic advantages. Conversely, ethical gaps between allied militaries could have dangerous consequences that imperil both political cohesion and coalition success. More specifically, if allies do not agree on their responsibilities and risk analyses around military AI, then gaps could emerge in political willingness to share risk in coalition operations and authorization to operate alongside one another.

Even though the United States is the only country to have *adopted* ethical principles for defense, key allies are formulating their own frameworks to account for ethical risks along the AI lifecycle. This

report explores these various documents, which have thus far been understudied, at least in tandem. Overall, the analysis highlights both convergences in ethical approaches to military AI and burgeoning differences that could turn into political or operational liabilities.

The key takeaways are as follows:

- DOD remains the leader in developing an approach to ethical AI for defense. This first-mover position situates the JAIC well to lead international engagements on responsible military AI.

- Allies fall on a spectrum from articulated (France, Australia), to emerging (the U.K., Canada), to nascent (Germany, the Netherlands) views on ethical and responsible AI in defense. These are flexible categories that reflect the availability of public documents.

- Multilateral institutions also influence how countries perceive and implement AI ethics in defense. NATO and JAIC's AI Partnership for Defense (PfD) are important venues pursuing responsible military AI agendas, while the European Union and Five Eyes have relevant, but relatively less defined, roles.

- Areas of convergence among allies' views of ethics in military AI include the need to comply with existing ethical and legal frameworks, maintain human centricity, identify ethical risks in the design phase, and implement technical measures over the course of the AI lifecycle to mitigate that risk.

- There are fewer areas of divergence, which primarily pertain to the ways that allies import select civilian components of AI accountability and trust into their defense frameworks. These should be tracked to ensure they do not imperil future political cohesion and coalition success.

- Pathways for leveraging shared views and minimizing the possibility that divergence will cause problems include using multilateral formats to align views on ethics, safety, security, and normative aspects.

In analyzing allies' approaches to responsible military AI, this issue brief identifies opportunities where DOD can encourage coherence by helping allied ministries formulate their views, and simultaneously learn from other approaches to responsible military AI as part of its own RAI implementation efforts.

Table of Contents

Introduction

In February 2020, the Department of Defense adopted five principles for safe and ethical AI, building on the existing legal and ethical framework that supports AI that is responsible, equitable, traceable, reliable, and governable.[2] DOD is now implementing these principles, including in a forthcoming DOD Responsible AI (RAI) Strategy and Implementation Pathway as directed by Deputy Secretary of Defense Kathleen Hicks. International engagement features as part of this effort, and the United States is most likely to succeed in leading coalitions of democracies committed to ethical and responsible AI if it appreciates the similarities and the differences in how like-minded states conceive of safe and ethical AI for defense.

There is strong consensus that cooperation with allies is key to accomplishing U.S. goals in AI, including in, but not limited to, the military realm. How allies approach ethical and responsible AI in defense therefore has implications for the success of international defense cooperation, as well as the U.S. strategic position in AI-related competition. Already, allies recognize U.S. leadership in the DOD's first-mover approach to safe and ethical AI. To varying degrees, many are trying to emulate or differentiate their own approaches from the DOD principles. Understanding where allies are in their approaches to ethical and responsible AI in defense can help the United States fine-tune its international engagement. In this engagement, leadership in ethical and responsible AI in defense can be seen as a two-way street. Leadership implies understanding and working with other democracies considering the role of ethics and safety in responsible adoption, as well as drawing lessons from other countries that are implementing responsible AI in the defense sector. By engaging in both lanes, the United States can encourage convergence in allied approaches to responsible and ethical AI in defense and, in doing so, open the door for collaboration on AI innovation and implementation.

At the same time, significant differences in ethical approaches to AI in defense could imperil political cohesion and undermine coalition success. Politically, alignment on ethics is important because shared values are at the foundation of U.S. alliances.[3] This also

trickles down to the operational level, where differing views on ethics could mean that allies field their systems with different legal authorizations and rules of engagement.[4] If coalition partners deem each others' capabilities to be based on different legal, ethical, and doctrinal assumptions, then forces may not be able to communicate and operate together.[5] Further, if different ethical bases for capability development mean that some countries have higher thresholds for what they develop and contribute to coalition operations, then others may perceive them as not equally sharing risks to life.[6] As such, political cohesion and policy considerations about ethics could directly influence operational effectiveness. In other words, failure to align allied perspectives on AI ethics in defense will inevitably undermine the ability of allied forces to understand each other and work together.[7]

In this light, this issue brief seeks to provide policymakers and analysts with one view on how similarities between allied perspectives on ethical AI for defense create opportunity for increased collaboration, and how the differences that are beginning to take shape can undermine said collaboration. Alignment and collaboration start with an understanding of variations in definitions of terms like trustworthy AI, ethical AI, and responsible AI in the defense context. These definitions are often fluid, depending on the legal, ethical, and cultural traditions of different countries. But different conceptions of responsible military AI nevertheless share foundations that help frame the analysis here. Broadly speaking, this issue brief focuses on how defense stakeholders steward AI innovation and integration in ways that: (1) respect the moral and ethical reasoning that underpins the responsible use of force; (2) meet and enhance compliance with law, which is based on ethics and translates reasoning into concrete obligations; and (3) minimize risks and unintended consequences for a safer and more secure international security environment.

To uphold ethical, legal, and safe foundations of AI development and deployment in defense, allies coalesce around two shared themes in their approaches to responsible military AI. First is that decisions around the design, development, deployment, and diffusion of AI do not enter into a vacuum, but rather into an

existing, multi-layered legal framework. It is not controversial for democratic countries to declare their shared obligation to respect law in order to remain accountable.[8] This accountability is owed to domestic citizenries, to the armed forces themselves, and to allies and partners in coalition settings, as well as to adversaries and the international community at large. As such, for some allies, emerging conceptions of responsible AI are closely interlinked with responsible state behavior, with continued legal compliance as the minimum requirement.[9]

The second commonality in all of the frameworks examined here is a shared focus on human centricity. There are several definitions of human-centric AI, but for the purpose of this analysis, it can be understood as the idea that AI is designed to meet human needs and improve upon the role of the human.[10] Not all frameworks use the term itself, but all stress the central role of humans in that machines should not replace humans and that humans remain responsible and accountable for decisions. By extension, this means a common approach to designing AI systems in such a way that the human user is not expected to adjust her or his own decision-making capacities to conform to the technology.[11] The inverse would place the machine at the center of decision-making systems. Meanwhile, countries consider humans to be central to defense planning and operations, and stress in their positions that they do not think it moral or lawful to delegate human responsibility to machines.

This legal and human-centric framing informs allies' views on key questions related not only to responsibility, but also explainability, trust, and related concepts. These commonalities should be seen as a baseline for responsible democratic governance of military AI, which leaves room for nuance in how each country interprets and prioritizes these types of principles. These nuances are important because defense stakeholders in allied countries do not necessarily emphasize the same principles in their evolving approaches to military AI.

Before proceeding with the remainder of the report, it is worth stating that this study does not address the adjacent field of autonomy in weapons. While autonomous weapons undoubtedly

pose important questions about ethics and legality, AI and autonomy in weapons are interrelated topics that deserve to be treated independent of one another.[12] As described in relation to country perspectives, there are several ethical risks that AI systems can pose without entering into an autonomous system, and without figuring into questions of lethality. This analysis focuses on the types of ethical risks associated with intelligent systems such as mission support software, select command and control (C2) systems, cyber detection systems, enterprise AI, and intelligence-related systems, among others.[13] These types of AI systems could certainly be integrated into autonomous systems, but the associated ethical questions remain discrete.

Another reason the analysis focuses on AI and not autonomy in weapons systems is that the topic is already well covered in other literature.[14] The United Nations Convention on Certain Conventional Weapons has focused on lethal autonomous weapon systems (LAWS) since 2013, and this has been the primary format for technical expertise, civil society engagement, and diplomatic engagement to converge.[15] As a result, questions about ethics and legality of autonomy in weapons, and specifically LAWS, often involve diplomatic actors at the fore of domestic government approaches. Concerns about the ethics of intelligent systems, on the other hand, currently receive less attention in military debates. To maintain a stricter focus on AI rather than autonomy in weapons systems, this study focuses more on technical and policy approaches in defense ministries, which have more agency in ethical and responsible AI policy.

Still, bifurcating approaches to intelligent versus autonomous weapons is easier said than done because many countries choose to combine the two topics under a "military AI" umbrella. In this regard, it is worth noting that the United States has a different starting point from most of its allies, in that the 2012 DOD Directive 3000.09 on Autonomy in Weapon Systems established early policy and guidelines on autonomous and semi-autonomous weapons. For DOD, this facilitates different, complementary policy tracks for implementation of AI principles on the one hand, and policy and guidelines for autonomy in weapons on the other. Although the same may not be true for other allies' approaches to

military AI, the focus here on intelligent systems aims to facilitate a more direct comparison with the U.S. military conception of RAI.

A related caveat is that some allies are either in the midst of constructing, or choosing to not publicize, their approaches to ethical and responsible AI. As such, allies' formulations of responsible military AI should be seen as evolving processes. The narrow availability of information on AI ethics beyond autonomy in weapons may reflect political sensitivities, including cultural differences around how transparent defense ministries and armed forces are. As such, the hope is to fill this gap by addressing how allies conceive of AI ethics for defense.

With these notes in mind, the remainder of the report proceeds as follows: First, a brief section covers how DOD has approached safe and ethical AI in defense, highlighting in particular how international engagement features in its efforts. Next, the allies that are formulating public perspectives on responsible military AI are discussed in three categories: those with articulated views, emerging views, and nascent views. These categories fall on a subjective, flexible spectrum based on the breadth and depth of publicly available information. Articulated views—from France and Australia—are not necessarily officially adopted policies, but are the most elaborate in terms of what information is public. Canada and the U.K. are classified as allies with emerging views because of clear indications that they have ethical assessment frameworks and processes, but with less comprehensive documents available at the time of writing. Countries classified as having nascent views—the Netherlands and Germany—show some evidence that they are focusing on responsible military AI to differing degrees, and their views may be more prominent in multilateral formats. Because this structure is based on availability of information, other key allies like Japan and South Korea, which have not issued public views, are not included in this analysis.[16] Select multilateral perspectives are also important complements to the national-level discussions, and are included thereafter. NATO and the Joint Artificial Intelligence Center (JAIC)-led AI Partnership for Defense (PfD) are discussed as formats that already have established processes to collaborate on responsible and ethical AI in defense. The European Union (EU) and Five Eyes are also selected as cases

for this analysis given their scope to potentially issue more comprehensive, public approaches in this policy area. Prior to concluding, a brief section on implications seeks to consolidate the findings from the previous sections and extrapolate main similarities, differences, and possibilities for more cohesive approaches to ethical and responsible military AI.

In doing so, the goal here is not to parse semantic differences between principles themselves, but rather to clarify how allied approaches respectively align with and differ from the U.S. position in their implementation pathways.

Brief Overview of the U.S. Approach to Ethical and Responsible AI in Defense

Main documents from DOD include:

- U.S. DOD AI Strategy (2018);
- Defense Innovation Board's (DIB) AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense (2019) – primary and supporting documents;
- Memorandum on Implementing Responsible Artificial Intelligence in the Department of Defense (2021).

Following the adoption of the DOD AI Strategy in 2018, the U.S. approach to AI ethics in the defense realm can be generally broken down into three phases: (1) the DIB leading the process to define AI ethics principles, (2) DOD adopting these principles for safe and ethical AI, and most recently, (3) the beginning of more visible efforts to implement RAI across the Department and armed forces.

Starting in July 2018, the DIB began its 15-month process on safe and ethical AI for defense, with the mandate of recommending principles to DOD in its capacity as an independent federal advisory committee.[17] This process took the form of public consultations, listening sessions, the formation of an informal DOD Principles and Ethics Working Group, expert roundtables, a

classified “red team” session, and a tabletop exercise.[18] As part of these consultations, government officials from “close partner nations” were also involved—including as part of the monthly meetings of the informal DOD Principles and Ethics Working Group.[19]

The role of allies in the resulting DIB recommendations largely focuses on the intersection between AI ethics and international norm development. More specifically, the DIB conceived of the role of allies mainly through the lens of DOD leadership, focusing on “how AI will be developed and used, and whether there ought to be any regulation on particular applications” to mitigate potential harms.[20] This is seen hand-in-hand with DOD’s “duty to the American people and its allies to preserve its strategic and technological advantage over competitors and adversaries who would use AI for purposes inconsistent with the Department’s values.”[21] In other words, the DIB sees aligning technological development, deployment, and intended outcomes with democratically informed values as a strategic obligation just as much as a departure point for the U.S. to lead norm development in the international community.

The culmination of the 15-month process came in October 2019, when the DIB articulated its five recommended principles for safe and ethical AI. The five principles that DOD then adopted in February 2020 are largely similar to those that resulted from the DIB-led process:[22]

- “**Responsible**. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities.

- “**Equitable**. The Department will take deliberate steps to minimize unintended bias in AI capabilities.

- “**Traceable**. The Department’s AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable

methodologies, data sources, and design procedure and documentation.

- "**Reliable**. The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles.

- "**Governable**. The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior."[23]

Since DOD adopted these five principles, the JAIC has led their implementation both in staffing and in processes. Further, implementation has also included efforts related to "procurement guidance, technological safeguards, organizational controls, risk mitigation strategies and training measures."[24] On training measures in particular, the JAIC organized a RAI Champions Pilot to educate multidisciplinary military AI stakeholders on AI ethics and implementation.[25] The eventual development and implementation of "governance standards" that encompass these measures, as included in the responsibilities of the JAIC Head of AI Ethics Policy, are namely geared toward internal use.[26] Further, such governance standards can also guide alignment efforts with allies and partners—as then-JAIC Director Lieutenant General Jack Shanahan mentioned with regards to using ethical principles to "[forge] a path to increase dialogue and cooperation abroad to include the goal of advancing interoperability."[27]

These priorities are also seen in the JAIC's international engagement. The JAIC is focused on "shaping norms around democratic values" as one if its three pillars of international engagement.[28] The other pillars of international military AI policy—"ensuring data interoperability and working to create pipelines to enable the secure transfer of technology"—also partially depend on ethics, safety, principles, and possibly even regulations.[29] Importantly, some technical aspects of this engagement concerns

adoption issues that are not discussed at length here. Nevertheless, when taken together, these three pillars not only refine the DIB-recommended goal of DOD leading international norm development, but also provide scope to align implementation with allies and partners—as is picked up in the section on the PfD below.

In May 2021, the Biden administration reaffirmed the dedication to DOD's ethical principles and directed new actions for implementation with a focus on RAI.[30] In her memorandum on *Implementing Responsible Artificial Intelligence in the Department of Defense*, Deputy Secretary of Defense Kathleen Hicks announced the formation of a senior-level RAI Working Council to "accelerate the adoption and implementation of RAI across the department."[31] Following its training based on the RAI Champions Pilot, the RAI Working Council will work closely with the JAIC on various issues that are referenced in this report.

To facilitate comparisons with allies' approaches to AI ethics for defense, two taskings of the RAI Working Council are particularly relevant. The first is the broad, overall aim of leading responsible AI globally. To this end, creating a "Responsible AI Ecosystem" is one of the foundational tenets of RAI implementation.[32] Here, international components of this ecosystem include allies and partners to enable better multi-stakeholder collaboration and to also advance norm development "grounded in shared values."[33] With the overall aim of leading responsible AI globally, this foundational tenet is the only one that explicitly names international engagement. Yet other foundational tenets are also relevant to assess convergences and divergences, including: governance structures for accountability; trust based on testing, evaluation, validation, and verification (TEVV); and a whole-of-lifecycle approach to risk management.[34]

The second relevant tasking is that the RAI Working Council is due to submit a report to the Deputy Secretary of Defense by late September on "policy modifications to enable RAI considerations within existing supply chain risk management practices."[35] In particular, the introduction of supply chain concerns entails a broader focus than the DIB initially considered in its

recommendations. The DIB did look at traceability of data sources, and DOD has also separately considered how reliance on foreign-sourced parts may impact the resilience of its systems and ability to maintain a technological edge.[36] Yet the connection between RAI and supply chain risks is a newer question in U.S. documentation on ethical and responsible AI pathways. As described below, this is not necessarily the case for allies.

From the DIB process through RAI implementation, these three phases show coherence in the U.S. approach to AI ethics in defense. Allies and key partners feature prominently across these three phrases, primarily as part of norm development and a shared policy basis for aligned, interoperable forces.

Articulated National Views: France and Australia

France and Australia are the main key allies whose publicly articulated approaches to ethical AI for defense are advanced enough that they can be assessed relative to the DOD principles. The views are not necessarily adopted government positions, but key documents from the French Ministry of Armed Forces and Australian Department of Defence still reveal useful comparisons to the official U.S. position.

France

Main documents from the Ministry of Armed Forces include:

- AI Task Force's Artificial Intelligence in Support of Defence (2019);

- Defence Ethics Committee's Opinion on the Augmented Soldier (2020) and Opinion on the Integration of Autonomy into Lethal Weapon Systems (2021).

The French Ministry of Armed Forces has been studying AI-associated risks since 2019.[37] That September, France became the first (and, still at the time of writing, only) European ally to publicly issue a dedicated military AI strategy.[38] Ethics and responsibility appear throughout the strategy, most notably as aspects of "controlled AI" and in the announcement to establish a ministerial

Defence Ethics Committee.[39] Subsequently, as described below, the advisory opinions of the Defence Ethics Committee also lend insights into French views on responsibility and other concepts. Each of these building blocks is indicative of French thought and implementation pathways for ethical and responsible military AI, even if not in the form of adopted principles.

The French military AI strategy, called *Artificial Intelligence in Support of Defence*, describes "controlled AI" as the overarching framework for Ministry guidelines on AI adoption, including aspects that relate to ethics.[40] Notably, control (*maîtrise*) in this context refers to harnessing and governing AI, and is not synonymous with human control (see Appendix II).[41] In addition to adoption priorities like the imperative to maintain freedom of action and interoperability with allies, the guidelines see "trustworthy, controlled, and responsible AI" as interlinked concepts under the headline "guidelines for controlled defence AI." These three concepts come to light in the need for the Ministry to "have robust and secure systems which can be trusted to assist service personnel and commanders, dispelling any 'black-box' effect, while retaining human responsibility for action."[42] The rest of the section unpacks what trustworthiness, control, and responsibility mean, as gleaned from both the French military AI strategy and the Defence Ethics Committee advisory opinions on other technologies.[43]

Trustworthy AI is linked to robustness and security because the French identify the potential risk that humans may *not* trust critical systems with opaque or inexplicable results.[44] More specifically, they see risks of erroneous results from AI systems stemming from unrepresentative (biased) training data, "malfunctioning algorithms," and "insufficient understanding" of system behavior. This leads to a technical approach to validating trustworthiness, wherein the Ministry decides on the level of trust and robustness necessary based on the function of the AI system at hand. This means that the ministry would conduct a risk analysis during the design phase to determine how critical the function is.[45] Then, according to the "criticality" designation, development of the system would need to meet a certain threshold of safety, security, and explainability in order to be validated.[46] Considerations about bias could also be taken into account, especially if they overlap with

deception techniques.[47] Appendix III shows how the Ministry intends to categorize functions and decide on their corresponding thresholds; broadly speaking, safety- and mission-critical systems may require higher thresholds of validation to be deployed. The French anticipate operating not just in unknown environments, but often in communications-denied ("degraded") ones too, meaning that establishing operator trust in AI requires the systems to behave predictably amid situations they haven't encountered before, including interference and deception.[48]

These operational specificities are heightened in the military realm. But the reliance on technical measures here echoes the national AI strategy, which similarly identifies explainability and the need to consider ethics from the design stage as important to ensure accountability and, by extension, social acceptability, for AI.[49] However, this approach is dependent on validation, and the military AI strategy does little to describe *how* current validation processes may be insufficient for AI systems that either perform differently in unknown contexts in which they are not validated, or whose capabilities change over their lifecycle.[50] The French answer is that AI standards and certification schemes are necessary, and that the Ministry of Armed Forces should be involved in such efforts. But without more detail, it is difficult to determine how trustworthiness is assessed after the design phase. This is where aspects of control come in, with the French military AI strategy establishing a relationship between controlled AI and auditability, to account for ethical concerns later in the AI lifecycle.

The French military AI strategy only implicitly deals with the ethical aspects of data, but nevertheless highlights the need for data sharing to comply with regulations, security requirements, and appropriate use.[51] More specifically, data governance includes documentation practices ("configuration") to keep track not only of which algorithms are used, but also their "learning elements, their combinations and their data" (*e.g.*, training data, weights, parameters, design procedures, annotations on limitations, etc.).[52] This means that the French see documentation as important not only to delimit the abilities of AI components in relation to their intended use, but also as an area where standardization can help humans exert control to be able to trace systems. Indeed, the

French see traceability and conservation of data as part of governance (especially for traceability of malfunctions).[53] Oftentimes, ethical AI frameworks tie explainability, transparency, traceability, auditability, and documentation together.

On this note, the French conception of controlled AI goes a step further and ties auditability to the core value of sovereignty in French strategic culture.[54] This is because the relationship between auditability and control stems from geopolitical concerns. The strategy states, "France cannot resign itself to being dependent on technologies over which it has no control. [...] Preserving digital sovereignty therefore also involves controlling the algorithms and their configuration, and the governance of data."[55] This need for control comes from a desire to exert independence from the "stranglehold on AI exerted by China and the United States," including by strengthening European cooperation.[56]

While the geopolitical aspects and prospects of France to assert this independence are beyond the scope of this study, it is notable that they trickle into the French approach to trace the provenance of models and data. In particular, weapons are "critical applications" that will need to be auditable.[57] If enforced, this means that questions about data rights and legal authorities to transfer data (including from foreign suppliers) could render AI "uncontrolled" per the French definition. Here, protectionism straddles the line of ethics and adoption, with digital sovereignty as a potential factor that determines acceptability of both. This can also be seen in the imperative to maintain "freedom of action."[58] Control and freedom of action may be seen as reinforcing concepts because they both relate to the need to maintain independence. Moreover, both concepts stress the responsibility of the state to ensure that humans are accountable for their use of technology.

In the French strategy, responsible AI refers to human responsibility for continued adherence to legal obligations and the tradition of military ethics. Though only implicitly defined, this human responsibility is expected to be retained across all processes and institutional structures, with an emphasis overwhelmingly on the use of force and chain of command.[59] The focus on human command ensuring that force is used responsibly

refers at least equally to AI and autonomous systems—the latter of which the French explored in greater depth in documents separate from the military AI strategy.[60] Most notably, as detailed below, the concept of responsibility is present in the advisory opinion documents of the ministerial Defence Ethics Committee.

The establishment of the Defence Ethics Committee was announced in the French military AI strategy to ensure a continued emphasis on ethics, and to issue advisory opinions on emerging technologies in defense. Since being established in January 2020, responsibility has featured as an important theme in its initial mandate covering two key areas of focus: augmented soldiers and autonomy in weapon systems.[61] While AI fits into both issue areas, it is notable that AI itself did not merit its own categorization. Indeed, the advisory opinion on autonomy in weapons reiterates that aspects of AI like non-lethal decision support systems are beyond its scope, and such issues are not yet part of the Committee's agenda.[62] Nevertheless, the resulting advisory opinions both connect to the concept of responsibility because both human augmentation and autonomy in weapons indicate the clear priority of maintaining human centricity in decision-making. Both opinions of the Defence Ethics Committee stress the need for humans to maintain their decision-making capacities not only from a moral standpoint, but also for the sake of operational continuity. In both cases, the French positions on ethics related to emerging technology are conscious of the risks of over-reliance on new technological enablers. Be it preventing addiction to human-augmenting medication and devices, or preventing automation bias in human-machine teaming, the Defence Ethics Committee makes clear that responsibility also requires humans to still be capable of achieving operational objectives even without assured access to the technology.[63] In short, responsibility means not only that humans should remain responsible and accountable for decisions, but if also considering the opinions of the Defence Ethics Committee on adjacent technologies, also that it would be irresponsible to excessively rely on technology.

For the French, another concomitant part of responsibility is how technology changes the relationship between the operator, combatants, and non-combatants in-theater. On this point, the

Defence Ethics Committee advisory opinion on autonomy in weapon systems also stresses that the distance of the operator from the operation itself can alter her or his judgment.[64] Such ethical tensions can have consequences for the foundations of responsibility in the French philosophical tradition. One question that senior French military officers have posed is whether it is unethical for two sides to face significantly uneven levels of risk, or whether military ethics require combatants on both sides to face risk.[65] To be sure, the operational advantage of reducing risks to the safety of one's own forces is imperative—and it could equally be argued that it would be unethical to *not* reduce risk on your own side when possible.[66] Without minimizing this obligation, it is worth noting that French ethics doctrine defines a soldier's own responsibility in part relative to the risk she or he faces. Per the Chief of Army's "Exercise of the Profession of Arms: Foundations and Principles," French soldiers derive legitimacy from the state, which confers the "responsibility to inflict destruction and death, at the risk of his life, in respect of the laws of the Republic, international law and the practices of war."[67] French military ethical debates around remote operators center on the phrase "at the risk of his life," questioning whether imbalanced risk continues to fulfil the criteria of legitimacy.[68] As such, technology is introducing new questions related to identity—which could have spillover effects into views on AI ethics and responsibility in the French Armed Forces.

This philosophical debate is implicit in the Defence Ethics Committee's opinions, and is relevant to adjacent questions about AI for two practical reasons. First is that French military officers see the need for guiding principles for technology that increases operational distance or can be applied to grey-zone activity. They note that ethics and law are well structured in rules of engagement, commander's intent, and compliance with legal frameworks for conflict. But because these often only apply *above* a certain threshold of hostility, guidance that allows the armed forces to maintain "ethics and moral strength" for other types of military activity is lacking.[69] This includes governance for technology that increases the distance between operators and operations, including in the information domain. As technology

broadens and accelerates changes to the operating environment, taking these ethical and moral considerations into account could necessitate new "deontological principles" on duty and obligation for the French Armed Forces.[70] The concept of distance from operations enters into advisory opinions, both in relation to automation bias and psychological effects. But even with the advisory opinions of the ministerial committee, such a framework is still missing.[71]

Second is that different allied perspectives on distance from operations could affect coalition operations if political differences widen. If guiding principles on distance from operations enter into rules of engagement or prompt questions about commanders' intent and responsibility, then different ethical bases for the use of new technologies in warfare could create political tensions. More specifically, if the more technologically advanced allies send more AI-enabled support and fewer troops as their contributions to coalition operations, some allies may perceive that others are not willing to equally share the burden of risks to life.[72] If not managed, sensitive issues that stem from different ethical risk calculations could decrease political cohesion.

In sum, while France has not strictly defined principles that it can adopt to ensure safe and ethical AI in defense, the Ministry of Armed Forces has dedicated attention to the issue both in its conception of "controlled AI" and the aspects of trustworthiness, control, and responsibility that it entails. Furthermore, while it is notable that the new Defence Ethics Committee has gone straight to questions related to human augmentation and autonomy in weapons, rather than AI ethics as its own category, the implications of these AI-adjacent technology areas show a consistent emphasis of responsibility as key to the French articulation of ethical AI in defense.

Australia

Main documents from the Department of Defence (not official government position) include:

- Defence Science and Technology Group's *A Method for Ethical AI in Defence* (2020) and the tools it establishes:
    - Ethical AI for Defence Checklist;
    - Ethical AI Risk Matrix;
    - Legal and Ethical Assurance Program Plan (LEAPP).

Australia emphasizes pragmatic approaches to ethical risk management over the declaration of principles. This approach primarily draws from the Australian Department of Defence's *Method for Ethical AI in Defence* (hereafter Method) technical report. While not a formally adopted view of the government, it establishes tools to assess ethical compliance that are currently under internal review and are already being trialed.[73] Even as an opinion, the Method is the clearest articulation of ethical AI for defense among the Indo-Pacific allies.

The Method document offers five "facets" for ethical AI in the Australian Department of Defence and armed forces, as well as an ethical assessment toolkit through which these facets can be operationalized. Overall, the five facets—responsibility, governance, trust, law, and traceability—echo the U.S. view that AI ethics for defense should focus on compliance with legal frameworks and moral obligations, as well as a functional approach to security and safety. While the Method is the result of a single workshop, the multi-stakeholder approach is similar to the longer DIB-led process in that it included experts from academia, industry, civil society, and bodies across the military and government.

Moreover, the focus of the Method is on pragmatic tools that can be used to implement ethical risk assessments. As such, these "facets" are not intended as adoptable principles to disseminate throughout the Department. As stated in the report, "rather than propose singular ethical AI principles for Defence, this report aims to provide those developing AI with facets of ethical AI that should be considered, including the questions to ask, topics to consider and methods that may be relevant to Defence AI projects and their stakeholders."[74] The focus therefore lies in implementation, as discussed below.

The Method's aim is to provide all relevant stakeholders in the AI pipeline with practical risk assessment tools that treat ethical risk on par with other types of risk such as safety and security. This includes one open-source tool, the Data Ethics Canvas developed by the Open Data Institute, and three original tools from the Method to "manage ethical risks."[75] These original tools are: an ethical assessment checklist; a risk matrix in Excel; and, lastly for projects deemed above a certain threshold of ethical risk, a formal documentation program called the Legal and Ethical Assurance Program Plan (LEAPP).[76] Together, these Australian ethical risk management tools are important not only to identify ethical risks, but also to follow up on them.

More specifically, the tools offer a process to validate that contractors have indeed taken the ethical risks they identified into account in their design and testing prior to later acquisition phases. In this way, they are procedural risk-assessment complements to technical validation and verification measures for the Australian Department of Defence. These tools are calibrated to the level of ethical risk that the anticipated use of the AI system would encounter or exacerbate. In addition to identifying the risks, procedural checks would validate that the contractors have followed through on addressing them—also with the incentive that it would be a worse outcome for them to find that unaccounted ethical and legal risks delay later stages of development or compromise the acquisition. The incorporation of ethics in design through the acquisition lifecycle also intends to build trust in the process and, by extension, the systems by the time they go into service.[77]

If high-risk projects and major weapon programs include AI components, then the risk assessment checklist would result in LEAPP being chosen as one of the ethical frameworks to comply with. LEAPP is a contractor's plan that seeks to give the Department of Defence "visibility into the contractor's legal and ethical planning; for progress and risk assessment purposes; and to provide input into the [government's] own planning."[78] For instance, LEAPP could be used as part of an Article 36 weapons review—a legal review of new weapons, means, and methods of warfare mandated for states party to the 1977 Additional Protocol

l to the 1949 Geneva Convention.[79] Article 36 reviews typically feature as part of legal discussions about autonomy in weapons, but nothing precludes them from use for non-autonomous, AI-enabled weapons as well. In this light, part of the LEAPP assessment would focus on technical measures that relate to international legal principles, regardless of the level of autonomy involved, which would be included in early negotiations between the contractor and the Australian Department of Defence.[80]

To be sure, compliance with the legal principles is non-negotiable. But in a more "iterative" fashion, the negotiations would focus on which requirements (*e.g.* software safety plan, safety management plans, human factors plans) would be necessary for the system to be certified for legal reviews, ethical requirements, as well as social responsibility more broadly.[81] As a result, this means that ethical risk would be identified in the early phases of design, and then carried through in TEVV. This negotiated framework between contractor and government is reserved for the higher-risk applications, and the connections it establishes between risk assessments and requirements validation is one of the most concrete practices that U.S. allies have thus far developed for AI ethics implementation in defense.

While the Australian focus is on tools rather than principles, the tenets in the Method are important bases for stakeholders to identify and assess ethical risks, including what frameworks AI developers consider in their own design processes. To this end, two aspects of the Australian tenets are worth briefly discussing.

First, for the trust tenet, the topic of "contestability" is unique to the Australian approach. The Method recommends importing the Australian government's civilian ethical AI principle of "contestability" into the military domain.[82] The Australian AI Ethics Principles define contestability as a "timely process to allow people to challenge the use or output of the AI system" that should be available for cases "when an AI system significantly impacts a person, community, group or environment."[83] Examples that apply both to the civilian or defense realms could include enterprise AI or recommender systems for promotions and human resources decision support. The aim of contestability is to ensure public trust

by allowing redress for harm, but the threshold of "significant impact" is fluid. As such, there is little guidance on what constitutes contestability.[84] If the Australian Department of Defence does adopt or enforce the contestability concept, however, then the high-risk nature of military activities could mean that the threshold of this harm is clearer than is the case in the civilian realm.

Overall, this intentional overlap between civilian and military is not unique to AI ethics tenets, and indeed builds on other Australian legal and cultural commitments to accountability. Legally, for example, there is no military court in the Australian Defence Force's disciplinary regime.[85] Culturally, the Australian "mateship" ethos bridges military history with national identity in a way that makes equality and respectful disagreement with authority—a basis for contestability—part of the Australian Defence Force's strategic culture.[86] In this way, building trust is not just about the technical measures like TEVV for reliability, but also about how individuals affected by AI can trust in the processes and structures responsible for its development and use.

Secondly, the lawful AI tenet is also worth highlighting because its explicit focus on *enhancing* compliance with international humanitarian legal principles is reflected in current Australian military AI consultation and collaboration with industry. The main point of lawful AI is that the technology is introduced into an existing legal and ethical framework—and that human-centered uses of AI should produce more ethical, and better humanitarian, outcomes.[87] In other words, lawful AI is not just about complying with the bare minimum of legal obligations, but also improving the standards of compliance already in place. The Australians point to two associated topics that lawful AI seeks to reinforce: "protected symbols and surrender" and "de-escalation."

On this point, the focus on protected symbols in lawful AI is also worth highlighting because it can currently be seen in Australian industry. Current Australian military AI development includes AI-enabled decision-support systems that improve compliance with international legal principles. Conceptually, this applies to what Australian researchers refer to as Minimally Just AI, or "MinAI," systems. Research on MinAI is not driven or mandated by the

government. Still, it corresponds to the results from the Defence-led ethical AI workshop in that it seeks to identify "protected symbols, protected locations, basic signs of surrender (including beacons), and potentially those that are *hors de combat"* so that human operators avoid them.[88] In practice, a small Australian company is currently using the Method as a guide for its system, called Athena AI, which identifies and classifies objects that "must not be targeted for legal or humanitarian reasons," including battlefield hospitals and other protected sites.[89] This includes validating data and designing features for end-users to understand the limits of the AI system, such as a notification if the computer vision system is less confident about a classification, so as to prevent cognitive bias.[90] Using the tools, scenario development, and a series of workshops on ethics and legality, Athena AI was designed in tandem with a 70-page legal and ethical framework.[91] As is noted below, the pragmatic tools from the Method could similarly inform Five Eyes activities.

Overall, the lawful AI tenet and focus on ethical risk assessment tools, rather than principles themselves, is a different packaging of the legal framework in which all allied democracies situate their views on ethical AI in defense. The Method contends that "lawful AI" has "no equivalent" principle in its comparison to other ethical AI frameworks.[92] In part, this is a superficial difference, as abidance to law is a preamble and is embedded in principles themselves, including those that DOD adopted. Signaling-wise, this more direct emphasis on law may intend to connect assessments of ethical risk with reputational risk. Risk management is just as much an entry point to ensure ethical AI efforts correspond with concrete, existing practices in the Australian Defence Force, as it is a form of responsible state behavior.[93] Contestability is one example of this—as a mechanism that seeks to enhance responsiveness to democratic citizenries and enhance social acceptability of AI—including here for defense. For military AI cooperation with Australia, it will also be important for allies such as the United States to know whether cooperative activities can be subject to formal contestation procedures.

Emerging National Views: the U.K. and Canada

While France and Australia have issued public approaches to their views on ethical AI as part of adoption, it is possible to deduce the approaches of the U.K. and Canada, both key allies, based on initiatives they have led on AI in defense, ethics, and data governance. The U.K. Ministry of Defence and Canadian Department of National Defence positions are considered emerging based on the availability of information in the public domain, which indicate strong foundations for more articulated approaches.

The U.K.

Main public documents from the Ministry of Defence include:

- Defence Science and Technology Laboratory Biscuit Book 2020: Building Blocks for AI and Autonomy;
- The announcement of a forthcoming U.K. Defence AI Strategy.

In the U.K., ethical and normative aspects of AI feature in recent strategic documents, including the government's Integrated Review of national security and international policy, and in the Ministry of Defence's accompanying Command Paper published a week later in March 2021. The Integrated Review names "supporting the effective and ethical adoption of AI and data technologies" and "identifying international opportunities to collaborate on AI R&D, ethics and regulation" as aspects that can help build public trust and early adoption of military AI.[94] This is consistent with the Ministry of Defence's contributions to achieving the British strategic interest of "the ethical development and deployment of technology based on democratic values," as reaffirmed in the Command Paper.[95] One area of daylight between the two documents, however, is the Integrated Review's concern about the gap between the pace of global governance and the development of standards and norms, in contrast to the Command Paper's stated need for "standards and norms for the responsible and ethical adoption of these new technologies."[96]

How exactly the U.K. Ministry of Defence will approach these interrelated military governance challenges is due to become clearer in the near future. More specifically, the U.K. plans to establish a new Defence AI Centre in order to centralize its AI developments.[97] Further, the U.K. Ministry of Defence is planning to publish a Defence AI Strategy that will incorporate ethical adoption considerations.[98] A ministerial AI ethics committee is also currently analyzing AI in defense, including issues related to trust.[99] In terms of oversight, both the new Defence AI Centre and this committee are important developments to bridge ethical AI endeavors at the working level with a higher degree of political and strategic attention.

The U.K. approach to military AI adoption includes a process for developing guidelines on ethical AI, which includes public-facing aspects led by the Defence Science and Technology Laboratory (Dstl).[100] Dstl established an AI Lab in 2018, which has made it the natural home for technical questions related to ethics, risk, and safety concerns.[101] While few details of the ministerial AI ethics committee are available at the time of writing, Dstl's activities advancing AI ethics in defense provide an indication of the U.K. approach. For instance, Dstl sponsors an ethics fellow at the Turing Institute to focus on "improving robustness, resilience, and responses of systems that support logistical, tactical and strategic operations, as well as wider applications in urban analytics, cybersecurity and social data science."[102] Furthermore, in 2020, they also hosted a conference that focused on safety, robustness, trustworthiness—which is part of the process on creating ethical guidelines for military adoption of AI.[103]

Notably, the Dstl AI Lab also debuted a "biscuit book"—a kind of introductory guide for contractors—that covers AI safety and ethics. Although predominantly a technical research organization, the building blocks in this book extend to non-technical considerations that should be factored into development of AI and autonomy in weapons. Two of the building blocks—consent and confidence—are relevant here and are described in more detail in Appendix V. While consent comprises ethics, risk and policy appetite, and legality, confidence includes more of the safety and security measures that are part of responsible military AI.[104]

Nevertheless, although the questions are a helpful guide intended for the Ministry of Defence customers, no mechanism currently obliges AI stakeholders to follow the guidance.[105]

As a last note, one question that the Defence AI Strategy may answer is the extent to which the U.K. considers ethics as important to its data governance efforts. One aspect that is less clear in the U.K. is how defense data governance accounts treats data ethics. So far, both in the biscuit book and more importantly the 2021 revision of the Defence Data Strategy, ethics does not feature as part of the imperative to govern data for aspects such as traceability and auditability. Furthermore, ethics is not included as part of the data governance architecture within the Ministry, which encompasses a defense information steering committee and specialist boards.[106] With this in mind, there are important questions to answer in the forthcoming Defence AI Strategy and more information from the ministerial ethics committee.

Nevertheless, when the U.K. publishes more information on its approach to ethical AI in defense, it will have strong national foundations to build on.

Canada

The main public document (not official government position) from Department of National Defence (DND) is:

- Defence Research and Development Canada's "Military Ethics Assessment Framework" in *Identifying Ethical Issues of Human Enhancement Technologies in the Military.*

Canada is in a similar position to the U.K. in that it has significant groundwork in place for its approach to ethical and responsible AI in defense. There is no publicly available Defence AI Strategy to analyze. Nevertheless, the Defence Research Development Canada Military Ethics Assessment Framework, as well as the DND guiding principles on ethical management of data, offer points of comparison to assess convergence and divergence with the U.S. approach.

The Military Ethics Assessment Framework can be seen as a technology-agnostic framework designed to be "broad enough to identify ethical challenges raised by any type of emerging technology."[107] This includes a strong legal framework, safety and security, accountability and liability, privacy, and trust, among others.[108]

As a technology-agnostic framework, the aim is for it to be used "purely as a risk assessment tool to help identify potential ethics questions that may be raised when considering the implications of technology use in the [future operating environment], developing policies surrounding technology implementation or preparing for challenges of encountering new technologies used by allies or adversaries."[109] Importantly, the possibility of disruption by adversaries is accounted for not only in security concerns as is the case for other allies, but also in "preparedness for adversaries" as its own category in the framework.[110] This general category ties the level of risk to expected operational realities. Overall, this Military Ethics Assessment Framework is intended to be used to *identify* when risk is heightened enough that policymakers should dedicate extra attention to it. Yet the Framework does little to advise *how* heightened risk should be managed. In this way, its practical use may be limited more to decide whether projects should be undertaken in the first place. When it comes to guidance for how ethical risk can be accounted for in development processes, procurement guidance, or legal assessments, it does less to offer solutions.

Canada is also working through data governance, an area that likely falls under the "privacy, confidentiality, and security" category in the Military Ethics Assessment Framework. To this end, data ethics are considered as part of the guiding principles of the DND Data Strategy principles, but with little clarity on their implementation. More specifically, there are guiding principles on managing data ethically "throughout their lifecycle to eliminate bias, ensure fitness for use, and adhere to the *Code of Ethics*." Security and trust feature as related concepts in the accompanying principles.[111] Although this approach declares ethical management as a lifecycle-long principle to abide by, the Defence Data Strategy only considers ethics for the *use* of data in its definition of the data

lifecycle. This gentle contradiction becomes more relevant when considering that the Strategy does not include a dedicated line of effort to implement this guiding principle. Further, ethics is considered part of data literacy—which ranks as low or medium in the implementation priorities, rather than a central feature of the governance framework.[112]

The DND Data Strategy represented the first effort by Canada to create a defense data governance framework, and still does focus on ethics-adjacent considerations related to aspects like traceability, data quality, accountability, and security. DND is expected to fit into broader Canadian government views on these issues as well—and has already encountered difficulty in the data governance realm for inadequately protecting privacy as part of an effort using AI to improve workplace diversity.[113] The usage of AI in human resources may indicate the difficulty of instituting a data governance framework—let alone one that explicitly calls out the ethical issues at hand.[114] As DND is subject to the Directive on Automated Decision-Making and Canadian Algorithmic Impact Assessment, incorporating values into data governance could become an increasingly important area for accountability.[115] In this way, the government's expectation of Defence abiding by the civilian framework is worth noting as a point of differentiation from the United States, whose leadership on ethical AI in defense took hold amid a less extensive background of civilian digital legislation.

All in all, these emerging approaches to ethical and responsible AI in the U.K. and Canadian defence contexts are promising starting points for more robust engagement.

Nascent National Views: Netherlands and Germany

Many other allies have not issued public approaches to responsible and ethical military AI. As is detailed below, these countries may establish their views in multilateral formats, rather than articulating a national approach. This includes countries like the Netherlands, whose approach thus far echoes those of allies mentioned above, as well as Germany, albeit in the adjacent context of multi-stakeholder AI governance for international security. Their interest could also lie more in international-security norms and rules in the

international technology order, and are therefore considered nascent here based on their comparability to the U.S. position.

Netherlands

Main documents from the Dutch defense sector include:

- The Netherlands Organisation for Applied Scientific Research (TNO) technical report Artificial Intelligence in the Context of National Security—Final Note Study within the National Security Analyst Network (2020);

- A potentially forthcoming defense AI roadmap from the Ministry of Defence.

The Netherlands is often cited as one of the European allies with a more fully fledged approach to AI in military affairs—including in the ethical domain. Indeed, the Dutch Ministry of Defence has sent a foreign exchange officer to the JAIC to work on collaborative approaches to responsible AI. A Dutch defense AI roadmap or "vision" document has been in the works for some time, as mentioned in the Dutch government's Strategic Action Plan for AI.[116] Parliamentarians have also recently inquired about the ethical and legal aspects of NATO's approach to emerging and disruptive technologies (EDTs).[117]

The Dutch defense research agency TNO also issued a technical report assessing experts and policymakers' view on AI risk in national security—including risks linked to control, trust, overdependence, and error.[118] The report is not directly comparable as it denominates risks, rather than coming up with a framework for stakeholders to mitigate those risks. For instance, security concerns around control being overridden, inaccuracy, errors, and interference do overlap with principles like reliability and governability, albeit in this slightly different context. Relatedly, even though it does not list exact risk factors that align with U.S. principles like responsibility, equitability, and traceability, this does not mean that the Dutch are not considering these principles more implicitly and in other contexts such as the defense AI roadmap. On the other hand, other risks that the TNO classify in the report do

overlap with the ethical approaches of other allies here—such as reputational risk for domestic and international accountability.[119] How these risks translate into ethical guidance for the Ministry of Defence and other national security stakeholders remains to be seen.[120]

Germany

The main document (non-governmental) is:

- Airbus-Fraunhofer Institute for Communication, Information Processing and Ergonomics (FKIE) joint "FCAS Forum" Working Group's *White Paper on The Responsible Use of Artificial Intelligence in FCAS – An Initial Assessment*.

The German approach to responsible military AI requires different considerations than the other allies surveyed here for two reasons. First, because the AI governance for international security policy space involves different primary actors. Second, because German political attention to autonomous weapons overshadows other aspects of military AI to an arguably greater degree than is true of the other allies surveyed here. German debates focus more on targeting—both for drone procurement as well as autonomous weapons development. The ban on LAWS, for example, is written into the government's coalition agreement.[121] The focus on autonomy in weapons is not matched by information about the ethics and safety of non-lethal, defensive applications of AI and AI-enabled decision support systems.

At the national level, it is the Federal Foreign Office, not the Federal Ministry of Defence, that has ownership over the policy space of AI governance and international security. The incorporation of the arms control agenda into military AI governance takes a broader definition of safe and ethical AI that incorporates questions related to diffusion and proliferation into account. This may be because arms control is a non-controversial topic that promotes German security policy without touching on more sensitive questions related to military adoption of technology.[122] As several analysts have noted, it is near-impossible to talk about military AI in Germany without the conversation immediately turning toward

LAWS.[123] Further, the focus on arms control lends itself to Germany's normative and diplomatic strengths. The framing of German contributions to responsible military AI in the realm of arms control is asynchronous—but not incompatible—with other countries' approaches to this policy area.

What this means is that the Federal Ministry of Defence is left with a backseat role.[124] This also heightens the stakes of multilateral efforts on responsible and ethical military AI, including for assessments of ethical risk stemming from issues like explainability or reliability. Indeed, Germany may be more active in these formats, especially in facilitating coordination between the EU and NATO given its longstanding interest in encouraging and facilitating EU-NATO cooperation.[125] Cooperation is already visible in other efforts related to technology and ethics—most notably in that the German Bundeswehr Defence Policy Office came up with views on future implications of human augmentation in collaboration with the U.K. Development, Concepts and Doctrine Centre.[126] The two countries share views on the future of operations, which may be productively channeled through activities related to policy alignment or, potentially, standardization.[127] Rather than going it alone, the German preference to cooperate—in bilateral and especially multilateral formats—may be seen as one way to focus on these issues with less domestic political pressure, and to substantiate contributions to defense partnerships.

Outside of government, responsible military AI initiatives are not necessarily on standby. For example, Airbus has partnered with the Fraunhofer FKIE research institute to ensure ethical compliance for AI and autonomy incorporated in the Future Combat Air System (FCAS) program.[128] FCAS is a cooperative project between Germany, France, and Spain to co-develop what they call a sixth-generation manned-unmanned aerial "system of systems." The Airbus-Fraunhofer FKIE partnership, dubbed the FCAS Forum, is supposed to "guide the development phase of the FCAS project from an overall societal and explicitly ethical perspective" to provide designers and developers with "ethical requirements and a process model for their implementation."[129] The engineers involved have alternatively described this as "ethical and legal compliance by design," which includes not only legal compliance, but also

considerations of "social acceptance."[130] Representatives from the two organizations see "technical implementation" as necessary to uphold the framework they develop.

To complement the technical aspects of responsible military AI, the FCAS Forum also includes a multidisciplinary expert panel that focuses on related aspects to responsible use.[131] While the panel includes representatives from the Federal Foreign Office and Federal Ministry of Defence, it is not a government-driven process. Further, at present, the expert panel is entirely German, and it is yet unclear if this industry-led initiative will echo the government's preference for cooperation by including French or Spanish subject-matter experts in the future.

One effort that the FCAS Forum has undertaken is a study on how civilian EU tools for AI ethics apply to the FCAS system-of-systems program. Specifically, the White Paper entitled *The Responsible Use of Artificial Intelligence in FCAS—An Initial Assessment* honed in on the EU Assessment List for Trustworthy AI (ALTAI) methodology in its attempt to identify frameworks that could encompass the societal and ethical implications of FCAS development.[132] Although EU AI policy, including this methodology, does not apply to high-risk designations in the defense realm, Airbus engineers involved in FCAS write that the ALTAI methodology "raises questions which are mostly applicable for FCAS and demonstrates consequences for the design and the operation of the system."[133] In other words, while the ALTAI methodology may not address uniquely military ethical concerns that weapon systems introduce, they find it nevertheless offers a starting point for a framework that defense stakeholders can consider.[134]

One question that the White Paper opened was whether the EU ALTAI methodology should be "extended or tailored to defence applications."[135] Though an answer to this question is beyond the scope of the report, the preliminary conclusions are that introducing the ALTAI methodology to operational and engineering teams on a case-by-case basis may help determine whether and how the methodology can be tailored to defence applications.[136] Further, they recommend that there should be training to

implement ethical design recommendations for relevant specialists.[137] They also suggested that future work could assess other requirements of trustworthy AI—as is partially addressed in Appendix VIII in this study.

To apply the EU ALTAI methodology to FCAS, the engineers first identified which "major AI case studies" would include their own discrete questions on responsibility and ethics. Their White Paper identified eight: mission planning and execution; target detection, recognition, and identification; situational awareness; flight guidance, navigation, and control; threat assessment and aiming analysis; cyber security and resilience; operator training; and reduced lifecycle cost.[138] Of these, the Airbus engineers deemed the latter three as "relatively uncritical from an ethical point of view."[139] If compared directly to the U.S. approach, there are several concerns that the DOD would deem relevant—or indeed critical—to safe and ethical AI. In particular, security and resilience are typically qualities that help build trustworthiness and reliability in AI systems. The FCAS White Paper *does* acknowledge that there are risks related to bias, deception, and the possibility of "AI-generated analyses and inferences [gaining more authority] in political decision-making."[140] Yet these are not connected to the aforementioned "uncritical" cases. Further, compared against other articulated guidelines and frameworks explored in this report, the approach to determining what *is* "critical" differs in that it does not necessarily calibrate ethical-risk mitigation measures according to the anticipated use of the technology, or the technique at hand.

Overall, the absence of concrete German, European, or transatlantic military AI frameworks means that industry has a different starting point when determining the most appropriate framework for responsible military AI. There is no immediately available information on government-guided implementation and requirements validation. With the government more focused on arms control, the German interest in a whole-of-lifecycle approach to AI governance may turn into a delegation of labor—with government looking at responsible use and diffusion and industry focusing on development. The mix here of both self-regulation and waiting for multilateral guidance indicates a clear German interest in military AI governance and ethics—even if more narrow

questions around autonomy in weapon systems continues to monopolize public debate.

This focus on autonomy in weapons is an important factor in assessing the degree of coherence between U.S. and German approaches to military AI, as it risks overwhelming less controversial issues. Beyond FCAS, the limited bandwidth for AI ethics beyond the tip of the spear could also mean the countries have different bases for how they develop and procure defensive systems and countermeasures. Further, it is not yet clear how the German military government is looking at AI ethics that are separate from autonomy in weapons. Without this separation, it could be more difficult to coalesce on views like the importance of cybersecurity and operator training to building trustworthy, reliable AI for defense.

At the same time, the German government's policy attention to AI governance and international security should not be discounted when it comes to norm development. As the JAIC's international engagement focuses on norms for a favorable technology order, there is room for the United States to learn from the German emphasis on arms control. In this vein, norms around the prevention of undesirable diffusion and confidence-building measures can be seen as part of responsible military AI.[141] Germany is already focusing on these areas, seeing multistakeholder engagement as necessary to modernizing 21st-century arms control beyond a narrow definition of state-to-state treaties. With industry already attempting to forge its own path on its ethical obligations, the Airbus-driven process could be seen as a complementary attempt to multi-stakeholder responsible AI in defense. But whether this delegation of labor on responsible AI to the private sector is a delegation of responsibility will depend on multilateral formats.

Multilateral Institutions Focusing on Ethical and Responsible AI in Defense

Understanding ethics and legality as part of the adoption of emerging technologies is not only a priority for democratic countries, but also a topic of interest for multilateral institutions

that are part of the security and defense architecture. Autonomy in weapons has been on the agenda at the UN level for longer than the countries highlighted here have spent time bridging technical and policy approaches to responsible AI in defense. Select multilateral institutions are also critical for consultations and alignment about these issues.

Beyond those mentioned here, several smaller allies may be waiting for multilateral views to then drive their own approaches to RAI, rather than dedicating resources to first issuing national views that will later have to align with broader multilateral structures. NATO is an obvious player in this domain, for reasons that are described below. The emerging PfD is important as an AI-specific multilateral format for like-minded countries—including non-treaty partners—to coalesce on this policy area.

North Atlantic Treaty Organization (NATO)

NATO is an important actor because it can help coordinate and facilitate consultations between allies to come to agreement on how ethical and responsible AI developments impact interoperability, cohesion, and operations. Further, the NATO Defence Planning Process is the primary defense planning tool for many Allies.[142] The focus on principles for "responsible use" is consistent with NATO's added value without dwelling too much on development—which happens primarily at the national (or bi-/multilateral levels outside of NATO).[143] Here, responsibility refers both to best practices in engineering (*e.g.*, ethical design) *and* responsible state behavior.[144] The North Atlantic Council and Military Committee—the senior civilian and military decision-making bodies in NATO—began contending with EDTs, including AI, in 2018.[145] This high-level interest built on several years of military and scientific experience at the working levels, as well as the introduction of conceptual and operational considerations in the workstream of Allied Command Transformation and the NATO Science & Technology Organization in the 2010s.

When presenting on EDTs to the senior civilian and military leadership at NATO in 2018, then-Supreme Allied Commander Transformation General Denis Mercier stressed that legal, ethical,

and political differences between Allies could "endanger our capacity to operate together."[146] He also focused on considerations around the "level of confidence in new technologies" as an adoption factor that is not purely technical.[147] This set the tone for political alignment on ethical concerns across the Alliance—already building on the foundations of the shared value embodied in the North Atlantic Charter and legal framework in which NATO operates. Subsequently, in October 2019, the Allies agreed to an EDT Roadmap that cited "legal and ethical norms" and "arms control aspects" as key technology areas among Alliance priorities to consider.[148] The political will to cooperate on technologies was solidified in February 2021, when NATO Defence Ministers endorsed an EDT Strategy.[149]

As part of the implementation of the EDT agenda, the NATO AI Strategy is expected to pick up on this theme in "guidance on both principles for responsible use of AI-enabled platforms and export control mechanisms."[150] Accountability and transparency—for weapon systems with varying levels of autonomy, as well as AI-enabled systems—and rules for industry may feature in this approach.[151] Already, the NATO Science & Technology Organization has identified a "strong emphasis on explainability, trust and human-AI collaboration" as well as "processes and standards for verification, validation and accreditation" as areas of interest for NATO, though they are not yet formalized in publicly pronounced principles.[152] Overall, the issuance of principles at the NATO level will be helpful to reflect priorities of multiple Allies, permit more Allies to align nascent or *ad hoc* initiatives to the broader agenda with one another, and potentially help bridge responsible AI with best practices or standardization that industry can follow.[153]

Technical reports that pre-date this EDT agenda also help establish a baseline for the Alliance's approach to responsible and ethical military AI. This includes NATO reports on human centricity and human control, as seen in Appendix IV.[154] The NATO Human View features in a technical report from 2010, and is notable as it commends a human-centric approach that has become popular in more recent civilian AI ethics frameworks (see Appendixes I and VIII). It is also an implicit theme found throughout the approaches

of all democratic countries who see the need to maintain accountability in human-defined frameworks to properly manage AI and the risks that can come with its development, use, and diffusion. The NATO Human View framework takes this a step further by developing a human-centric approach to network-enabled operations (*i.e.*, network-centric warfare) that ensures the human element is incorporated into capability development.[155] More specifically, NATO defines this human centricity by "depict[ing] how the human impacts performance (mission success, survivability, supportability, and cost) and how the human is impacted by system design and operational context (*e.g.*, personnel availability, skill demands, training requirements, workload, well-being)."[156] The Human View framework also applies the element to factors of human-system integration as well as "guidance on use of models to address uncertainty and/or discover emergent behaviors."[157]

NATO work on human control is largely focused on integrating weapon systems with different levels of autonomy, but still offers useful classifications that can apply to other EDTs like AI. As seen in Appendix IV, it establishes a framework of who the relevant stakeholders are, and what their corresponding roles are in delivering appropriate human control—a concept that could be adjusted to address other aspects of its forthcoming principles for responsible use. Some of these complexities are specific to autonomy and the use of force, as was defined in 2019. But it is also notable here that the NATO working group on "Human Systems Integration for Meaningful Human Control over AI-based Systems" was established in 2020. Importantly here, the working group suggests using the term "meaningful human control" outside of autonomy in weapons-related questions, which are primarily dealt with at the Convention on Certain Conventional Weapons. They suggest applying the term to address questions of accountability and human agency in a broader set of applications including intelligence, surveillance, and reconnaissance, planning, and decision support.[158] In addition to applying to these set of issues, the working group focuses on many similar aspects that also relate to responsible AI implementation, including: "national or organisational policy, systems specification, systems design,

systems [validation and verification], training of users, training of AI and [machine learning], systems of systems integration, C2 process development, interoperability, operational use, after-action review/lessons learned."[159] As such, meaningful human control may come to feature in responsible AI efforts within NATO.

While the issuance of principles and commitment to uphold shared values through responsible use are undoubtedly important, it is worth noting that NATO does have different considerations from nations (and militaries) as a supranational, non-regulatory body. Its more unique contribution could be standardization, as already incorporated into standardization agreements and training publications—the latter of which could focus on responsible AI in training. In anticipation of forthcoming guidelines or standards for industry, it is worth noting that NATO operational standardization also encompasses work implementing the Law of Armed Conflict into operational practice.[160]

For NATO, building on General Mercier's remarks that ethics and interoperability are linked, standardization on safe and ethical AI can likewise link to interoperability at the technical and procedural levels.[161]

AI Partnership for Defense (PfD)

In September 2020, the JAIC convened the inaugural PfD meeting, featuring virtual delegations from Australia, Canada, Denmark, Estonia, Finland, France, Israel, Japan, Norway, the Republic of Korea, Sweden, the U.K., and the United States to "shape what responsible AI looks like."[162] As of May 2021, three additional countries joined for the third PfD meeting: Germany, the Netherlands, and Singapore.

As the grouping of countries makes clear, the ability to include non-treaty allies in the PfD makes it a useful format to borrow from each other's approaches to RAI, be it to establish, refine, or implement nation-level views. Just two months before joining the PfD, for instance, Singapore prepared "preliminary guiding principles to be applied to the defence establishment in Singapore,

and Singapore's contributions to the global discussion on international norms for defence AI applications" in March 2021.[163]

Further, there may also be the possibility of taking aspects of responsible military AI from other countries that focus more on norms of responsible state behavior on board in the PfD format. Some allies explicitly mention a focus on norms, including the U.K. in its new national security and international policy, and Germany via its focus on arms control and emerging technologies. This normative emphasis harkens back to the U.S. approach to responsible and ethical AI in defense—which saw norms as one of the primary areas of engagement with like-minded countries. This normative focus could also benefit engagement with allies that have not yet begun any public iteration of views on responsible military AI, including Japan and South Korea. As such, the PfD's focus on responsible AI makes it an important venue to encompass technology norms that are based on democratic values and that focus on minimizing risks in the international security environment.

As a final note, it is not a coincidence that all allies surveyed here participate in the PfD. It is an important forum for them to exchange views—not only on aspects covered in this report, but potentially also the impact of civilian AI ethics frameworks and developments, as well as questions about autonomy-related aspects of human-machine teaming.

Other Opportunities for Multilateral Collaboration on Responsible and Ethical AI in Defense: the European Union (EU) and Five Eyes

Working with a number of multilateral institutions is critical to the United States' stewardship of AI aligned with democratic values and interests.[164] In addition to NATO and the PfD, the EU and Five Eyes are highlighted as relevant formats for cooperation on ethical and responsible military AI.

European Union

Of course, the EU is not an alliance—and the United States is not a member. But the EU's potential contributions to responsible

military AI are worth discussing here because of the implications of supranational EU policy on allies' own approaches to ethical and responsible AI in defense, as well as on EU-NATO cooperation.[165] Furthermore, the United States could also be more directly affected by EU policy as a third state that is eligible to receive R&D funding via the European Defence Fund. This is because the European Defence Fund, under the European Commission's authority, can require a mandatory ethics screening prior to fund disbursal.[166]

While the EU has adopted a bullish approach to trustworthy AI in the civilian realm, European institutions have been slower to define the implications for safe and ethical AI beyond the tip of the spear. Key civilian policies and regulations, like the General Data Protection Regulation and more recent legislation instituting the European approach to "trustworthy" AI, have clear carveouts for public safety, security, and defense. Still, the EU approach to civilian AI policy is relevant to transatlantic defense because the dual-use, general-purpose nature of AI means that military adoption of AI will depend on the ethical frameworks that dominate civilian development, regardless of carveouts. More directly, some European countries also choose to apply EU legislation like the General Data Protection Regulation to their own defense sectors, even though they are not required to do so.[167] With this overlap in mind, Appendix VIII overviews the applicability of the EU trustworthy AI principles for the defense realm.

In addition to examples such as Airbus' application of the ALTAI methodology to FCAS, it is notable that several European defense efforts mention the European Commission-supported guidelines for trustworthy AI as a positive step toward ensuring military uses of AI adhere to ethical standards. For example, in their co-authored food-for-thought paper on AI in defense, Finland, Estonia, France, Germany, and the Netherlands made explicit reference to the Trustworthy AI Principles, recognizing that the EU could leverage its normative power because of the centrality of ethical standards in AI for defense.[168] The focus on safety and security in EU AI policy also promotes "convergence between the AI community and the security community" to enhance robustness.[169] In sum, the emphasis on safety, security, and risk in EU AI policy is not only a

natural overlap, but also one that European defense stakeholders are seeking out.

However, it remains to be seen which EU body will take control of the responsible and ethical military AI agenda. There are various actors within the EU institutions that are largely beyond the scope of this paper.[170] Instead, there are only inklings of how the EU will approach responsible and ethical military AI at present. In the future, this topic could also feature in EU-U.S. security and defense dialogues.

For now, it is the European Parliament that plays the most visible role advancing ethics in European military R&D funding. This was seen in mid-2018, in its attempt to ban *all* military AI research using EU funds because of concerns about LAWS. The agreed-upon final version explicitly prohibits funding for LAWS at the European level—a deal-breaker without which the Parliament would never have agreed to allow for any defense funding.[171] But the final result was narrower because the EU does not have jurisdiction over its member states' armaments development unless they use EU funds. As such, while important, especially for dual-use and open-source systems, its jurisdiction on mandating ethical reviews is still limited and likely to not affect the majority of national, bilateral, or minilateral capability development programs. More recently, the Parliament-issued *Guidelines for military and non-military use of Artificial Intelligence* in January 2021 could indicate a stronger ethical bent than seen in the other institutions.[172]

Five Eyes

Including the United States, four of the Five Eyes countries are included in this report for their national-level approaches to ethical and responsible AI in defense.[173] Through policy exchanges and the Technical Cooperation Program (TTCP), the Five Eyes militaries are already engaged in cooperative digitalization efforts.[174] Policy exchanges between Australia, Canada, New Zealand, the U.K., and the United States take various forms that can facilitate alignment. TTCP is more specific as an "international organization aiming to collaborate and exchange defense scientific and technical research

and information, and harmonize and align defense research programs by sharing or exchanging research activities" between the five countries.[175] TTCP is directed by principals from the Five Eyes countries, who agree on and direct three-year Strategic Challenges on specific technical areas for collaboration. Here, the TTCP Strategic Challenges on cyber, autonomy, and AI are relevant here because they focus on aspects related to trustworthiness and related issues.

Most directly, the TTCP AI Strategic Challenge includes a Law and Ethics Working Group and also considers Trust and Transparency as one of its key themes.[176] The Trust and Transparency theme intends to better understand how agent transparency impacts human-system performance, and how to design the appropriate level of transparency in AI systems to increase trust in their intended use.[177] This lends itself well to responsible AI, for example, by applying ethical risk frameworks and tools from the different countries to cooperative activities. Already, the Australian Method includes recommendations to demonstrate "how the AI integrates with human operators to ensure effectiveness and ethical decision-making" in trials and exercises that simulate the "anticipated context of use."[178] Five Eyes countries have previously cooperated on areas that could link to this recommendation. In addition to sharing data for Project Maven, the TTCP Autonomy Strategic Challenge used the 2018 *Autonomous Warrior* exercise to trial the "Allied Impact" C2 software system focusing on human-machine teaming and integration of autonomous assets.[179] As part of the remit of the TTCP AI Strategic Challenge law and ethics working group, other frameworks discussed in this paper—namely the Canadian Military Ethics Assessment Framework, the U.K. approach, or the United States' responsible AI implementation measures—could equally follow the Australian recommendation to use collaborative activities like the 2018 *Autonomous Warrior* exercise to incorporate ethical risk frameworks into interoperability assessments.

TTCP collaboration on trustworthiness and technology also reaches back to the 2014 TTCP Cyber Strategic Challenge. Although it focused on cyber rather than AI, the four topics of its focus—vulnerability assessment, red teaming, building mixed levels

of trust systems, and developing metrics and measurements of trustworthy systems—offer a framework that responsible AI adoption could follow.[180] On the latter point, the TTCP working group developed an ontology-based assessment framework that formalizes metrics and attributes that constitute trustworthiness, which could inform allied approaches to trustworthy, reliable, and secure systems.[181] The Five Eyes researchers identified 13 core attributes, many of which dovetail with the broadly shared principles of ethical design and responsible use for AI: reliability; availability; safety; confidentiality; integrity; robustness; maintainability; adaptability; usability; timeliness; leanness; reactiveness; and proactiveness. Appendix IX details how these attributes are measured in relation to trust, resilience, and agility.

These attributes from cyber are relevant to AI because the TTCP group looked at factors that would affect the performance of the systems, including cyberattacks, human factors, and the impact of the physical environment.[182] For AI, equivalent concerns include protecting from failure modes, understanding how human interactions impact system performance, and validating and verifying the performance of AI systems in real-world conditions. Going forward, the trustworthiness framework that the TTCP Cyber Strategic Challenge established could also be used for the Five Eyes countries to understand how to use their allies' systems.[183]

Overall, because TTCP Strategic Challenges already combine technical research with experimentation, they are well-suited for implementation of responsible AI and ethical assessment frameworks.

Areas of Convergence and Divergence

For the most part, U.S. views on responsible and ethical AI for defense align with allies. Table 1 reviews the similarities, with the caveat that the absence of a similarity or equivalent principle should not be read as a point of divergence. As the country analyses above have shown, countries are at different stages of iterating and implementing ethical and responsible military AI.

In particular, Australia, Canada, France, the U.K., and the United States share responsibility and trustworthiness as core principles for ethical AI design, development, and deployment. While the United States does not use trust or trustworthiness as a standalone principle, it is embedded in all five of the DOD principles. This is similar to the U.S. conception of law as a cross-cutting theme that is implicit in responsible, equitable, traceable, reliable, governable AI. Other countries see trustworthiness as the overarching principle that comprises reliability, along with integrity and security. In this way, even though trust is not synonymous with reliability, it is the closest comparison. This is also affirmed in other countries' views: for instance, the Australian Method benchmarks its own facet of trust to the DOD principle of reliable AI.[184] These terms are interlinked and mutually reinforcing, as human-machine teams require operators to trust the systems they interact with. Focusing on reliability and security can help build that trustworthiness, especially to show operators that the systems are subject to rigorous testing and that processes are in place to appropriately calibrate trust depending on the capabilities of the system and the operating environment.[185]

Other principles and equivalent topics, such as controlled AI and feedback mechanisms for societal input, are also more explicitly laid out in allies' approaches to AI ethics. Here again, this is not to say that the United States does not share sovereignty concerns, but rather that these concerns are not as explicit in DOD's ethical AI principles as in allies' documents. Seeing national sovereignty as part of responsibility could come to be in tension with cooperation—as well as procurement decisions that breed dependence on the United States. It is notable that DOD is just beginning to insert supply chain considerations into its publicly available documentation on RAI. Meanwhile, it has been part of the French approach since they began considering ethical risks of AI in defense, and is also included in the Australian Method. As countries navigate this nexus, the extent to which sovereignty concerns fuel tensions between democratic allies will depend on other forms of cooperation.[186] Nevertheless, because of the overlap between security and assurance of control over the lifecycle of an AI system, responsible AI implementation pathways in the United

States may come to incorporate supply chain risks.[187] In this way, it would be similar to traceability and auditability concerns that countries such as France and Australia mention in their approaches to sovereignty in AI.

Additionally, differences between allies' views on responsible and ethical AI in defense may also stem from the extent to which other countries apply civilian AI policy and regulation frameworks to their own defense approaches. Although overviews of government principles and policies for ethical civilian AI are not discussed at length here, they become more visible in other countries, as well as the EU. Some of these aspects are encouraging—for example, German industry's voluntary compliance with trustworthy AI principles, to the extent they overlap with defense, due in part to the fact that there is no equivalent defense framework they can follow and implement themselves. Other entry points of civilian concepts into the defense realm include the Australian concept of contestability, Canadian DND subjectivity to the algorithmic impact assessment, and the choice of some EU countries to apply the General Data Protection Regulation to their defense sectors. These are both related to accountability as well as privacy, which are key differences that should be tracked even though allies overwhelmingly agree on the importance of ethics and safety for AI in defense.

Still, similarities between defense stakeholders include the view that militaries could not only inject new risks into their operating environments, but also expose their own organizations to risk if they leave ethical and legal questions unaddressed. Countries may have different ways to define and measure these ethical risks, as Table 1 implies, and as is detailed in the appendices. Overall, though, there is an implicitly common approach which recognizes that they must contend with the associated technical, legal, political, and moral risks from the front end of AI development. More concretely, they also agree that the way to implement responsible AI involves technical measures tied to safety and security, as well as procedures that make the legal context by which they abide as clear as possible.

Table 1: Comparison of allies' approaches to ethical AI principles and risks in defense

	Articulated views			Emerging views		Nascent views
	USA	**France**	**Australia**	**U.K.**	**Canada**	**Netherlands***
Responsible	Responsible	Responsible	Responsibility	Policy and risk appetite	Accountability and liability	Managing export/diffusion that enhances adversaries' capabilities
Equitable	Equitable	(Bias implicit in reliability)	(Part of trust)	Ethics (fair and equitable)	Equality	
Traceable	Traceable	(Mention of explainability/ configuration)	Traceability	Explainability	Privacy, confidentiality, and security	
Reliable	Reliable	Trustworthy	Trust	Trust; Assurance	Reliability and trust	Security (control being overridden; inaccurate decisions/error)
Governable	Governable (deactivate/ disengage)	(See control)	Governance	Resilience (fail gracefully)		System interference
Controlled		Controlled (sovereignty)	(Sovereign capability)			Level of dependence on foreign technology companies; control; lock-in
Lawful	Compliance with law	Compliance with law	Law	Legal	Compliance with law	
Societal feedback mechanisms			(Contestability)		Effect on society; consent	Public opinion; undue influence on the public

Key: blue squares are the primary principles/facets; white squares are related sub-topics mentioned in the respective documents

*The Dutch technical report focuses on risks, not ethics, and therefore has a slightly different scope than the others mentioned here.

Sources: Author's analysis of: U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," 2020; French Ministry of Armed Forces; "AI Strategy in Support of Defence," 2019; Devitt et al., "A Method for Ethical AI in Defence," 2020; U.K. Defence Science and Technology Laboratory, *Building Blocks for Artificial Intelligence and Autonomy, 2020;* Girling et al., *Identifying Ethical Issues of Human Enhancement Technologies in the Military* (2017); Neef and van Weerd, *Artificial Intelligence in the Context of National Security*, 2020.

How countries actually implement their shared technical and law-related priorities depends not on the principles or tools, so much as the oversight structures that are responsible for them. It is too early to assess differences in oversight structures, which will depend on the authority empowered to the ethics committees and AI-centric defense units that ministries are creating. Many of the documents overviewed here are technical reports or advisory opinions that are not binding in nature.

Whether these structures have mechanisms to mandate, or merely recommend, courses of action will also have implications for the extent to which they transpose their frameworks into action. As countries focus on implementation, their approaches to oversight may also emerge as differences between the structures working on ethics and responsibility of technology adoption.

Overall, pathways for leveraging shared views to advance the implementation of responsible AI should include learning from allies whose views emphasize both human responsibility and responsible state behavior extending beyond minimum legal obligations. Addressing these issues is not only a question of good engineering practices, but is also an exercise of responsible state behavior.[188] In a narrow sense, responsible AI can refer to ensuring that AI systems enter into human-centric frameworks that are defined by humans to maintain human agency and responsibility.

More broadly, though, it is notable that some allies see preserving freedom of action as part of a vision of responsible AI that encompasses responsible state behavior. Having a legitimate basis for military action is a feature of responsible state behavior, with civilian government oversight of militaries at its core to maintain accountability at home and abroad. This enters into the language of responsibility because operating in coalitions under multinational mandates can also confer international political legitimacy to operations.[189]

Allies with articulated views also translate their obligation to protect into language on military AI. This can be seen in arguments for the moral imperative to pursue AI-enabled capability development to maintain freedom of action and protect from

adversaries whose uses of AI do not respect legal and ethical obligations.[190] In these views, maintaining freedom of action can also mean maintaining interoperability, or even developing AI systems that help protect friendly forces, as most allies depend on cooperation to fill capability gaps.[191]

Beyond operational risks, responsibility also means incorporating ways to minimize risks in the international security environment. DOD has a role to play off the battlefield in this regard as well, including by developing norms around arms control. Allied concerns about diffusion and access, as embedded in the German international security and AI governance agenda, as well as risks that the Dutch identify, make this a compelling area for responsible AI cooperation between defense ministries.

In doing so, the United States could find complementary areas of interest with allies that see responsible military AI as encompassing norms in the international environment. In fact, this may be a palatable way to move debates beyond questions exclusive to autonomous weapon systems. While autonomy in weapon systems undoubtedly introduces important questions for ethics, legality, and responsibility, the dominant attention it receives tends to overshadow other aspects of military AI. This not only includes responsible AI implementation, but potentially even the responsibility states have to defend against AI-enabled threats from less ethical adversaries. The fact that most allies are still transitioning from ethical questions wrapped up in autonomy in weapons means that DOD can facilitate and complement their views on ethical and technical dimensions of AI in non-lethal or non-autonomous systems. In doing so, it could help steward the conversation toward other, underrated aspects of ethical design, development, and deployment of military technology.

With more of a focus on norms and state responsibility, a broader definition of responsibility beyond the DOD "responsible AI" principle can also introduce new convergences for U.S. RAI implementation. The DOD RAI Strategy and Implementation Pathway recently tasked by Deputy Secretary Kathleen Hicks can incorporate these views and help allies refine their approaches to AI ethics in order to enable greater cooperation and allied AI

adoption. This is because many allies see firm approaches to managing ethical risk as a prerequisite policy question before investing in AI-enabled capability development—including defensive systems and countermeasures.

Overall, international engagement is mutually beneficial to responsible AI endeavors. The United States should look at how other countries are implementing their approaches, just as the United States can exert influence and maintain its leadership role in responsible and ethical AI for defense by helping its allies and partners form their own views in alignment with one another.

Conclusion

No single actor has a monopoly on the answers to implementing responsible AI in any high-risk area, let alone defense. Cooperation is therefore important to collectively navigate the difficulties of responsible governance of emerging technologies. For DOD, the focus has been predominantly on the transposition of safe and ethical AI principles into action. Rather than adopting principles for defense, some allies are moving straight into implementation. Thus far, this is borne out in tabletop exercises, outreach, ministerial committees, ethical reviews, education and certifications, exercises and trials, and defense programs of record. It is too early to judge these fledgling efforts, but tracking their evolution may prove useful to broader AI ethics implementation, be it for other defense ministries or even civilian actors.

While jumping straight to implementation can mean a more pragmatic focus on tools, tracking how different AI stakeholders use those tools may be more difficult. In this way, principles can be seen as a helpful organizing force, as is the case for DOD. This said, the scale of the U.S. military bureaucracy and national security innovation base may require higher visibility relative to allied counterparts.

Still, another key difference is precisely this visibility. The analysis here is based on information in the public domain, which may also partially explain its transatlantic tilt. The U.S. approach to responsible and ethical AI for defense also differs from other

countries in that the consultation and process that led to its principles is far more transparent than is true for most allies. A possible Catch-22 could be at play here, with allies reticent to publicize approaches to such controversial issues, despite the fact that offering such inroads can build trust and confidence that governments are handling these high-stakes questions responsibly.

This is important not only for accountability, including to citizenries, but also because dedicating attention to responsible AI is a critical way to signal to industry, civil society, academia, and the research community that appropriate measures are not just boxes to tick, but are fundamentally embedded in the development of systems. In other words, responsible AI is important not just for public opinion, but also to strengthen relationships with the expert community that is rightfully concerned about the ethical implications of current AI advancement.

Further, the U.S. national security community should consider norms for diffusion and arms control as part of its responsible AI agenda in order to encourage more allies to issue public approaches to RAI. This does not necessarily mean leading the initiatives; indeed, if the next German government continues with the current arms control agenda, then following the German lead could become a more important area of cooperation.

While not all ethical questions relate to adoption, the overlap between them is important. To be sure, incorporating ethics into design prior to moving onto development and deployment phases is critical to ensuring a lifecycle view on AI ethics. But at the organizational level, it is important to make sure that military bureaucracies are neither engaging in "ethics washing" with no intention to implement anything, nor seeing ethics as their sole contribution as a way to "do" AI on the cheap. This is not to say that allies are not investing in AI-enabled capabilities—as many certainly are—but rather that these attempts and investments are so *ad hoc* that they could foreclose the possibility of scaling any such efforts. To this end, with the importance of ethics and legality for democratic accountability, tying ethical risk assessment

frameworks and associated processes with AI adoption could also be an inroad for more strategic approaches to AI integration.

This could be done through multilateral formats, with NATO and the PfD being the most appropriate for the time being. Also, using these formats to help allies dissociate the ethical and legal questions of autonomy from AI could push responsible AI in defense forward, including in associated technology areas not addressed in this paper, such as human-machine teaming or human enhancement. The fact that information about some allies' approaches to AI in defense can be answered via frameworks for human enhancement shows that the stakes of responsible AI are not just about adoption of this one, crucially important technology. The stakes are the testing grounds for the combinative technologies that lie ahead.

Author

Zoe Stanley-Lockman is an analyst researching military innovation, emerging technologies, and defense cooperation. The views expressed herein are the author's alone and do not reflect those of any organization.

Acknowledgments

The author would like to express appreciation of the CSET team for their support. Margarita Konaev dedicated time to this Issue Brief as if it were her own—and Lynne Weil, Matt Mahoney, Daniel Hague, Shelton Fitch, and Melissa Deng provided critical guidance and editorial support to improve its accuracy and readability. The idea to pursue this work would likely not have happened without Michael Raska's encouragement to begin studying ethics and governance of military AI in early 2019. Igor Mikolic-Torreira's subsequent support conceptualizing the project helped bring it to fruition. Thoughtful comments from David Whetham, Torben Schuetz, Maaike Verbruggen, and other experts who wish to remain anonymous capture nuances that enhance the comparability of the cases reviewed here.

Document Identifier: doi: 10.51593/20200092

Appendix I: U.S. conceptions of responsible AI

DOD	OECD Recommendation of the Council on AI	DIB expanded definition/context
Adopted definition: "DoD personnel will exercise appropriate levels of judgment and care, while remaining for the development, deployment, and use of AI capabilities" *Key terms for RAI Implementation as of May 2021:* foundational tenets of (1) responsible AI governance, (2) warfighter trust, (3) AI product and acquisition lifecycle, (4) requirements validation, (5) responsible AI ecosystem, and (6) AI workforce; actions for JAIC to coordinate, through Responsible AI Working Council, include (1) formation of Working Council and training; (2) DOD RAI Strategy & Implementation Pathway, (3) development of RAI workforce talent management framework; and (4) acquisition recommendations	"Responsible stewardship of trustworthy AI" as overarching framework for (1) human-centered values and fairness, (2) transparency and explainability, robustness, security and safety, and (3) accountability	Three layers of responsibility, with outer two layers focused on conduct of hostilities *First layer*: responsibility of persons with authorities over the design, requirements definition, development, acquisition, testing, evaluation, and training for any DOD system, including AI systems *Second layer:* responsibility mechanisms for action taken by decision makers during hostilities (Law of War, rules of engagement, commander's intent, doctrine)—including appropriate information on a system's behavior, relevant training, and intelligence and situational awareness at "epistemic thresholds" *Third layer:* remediation mechanisms after technologies have ended—both internal (accountability via Law of War and Uniform Code of Military Justice) and external to DOD (doctrine of State Responsibility)

(adopted – defense)	*(adopted by White House – civilian)*	*(recommended – defense)*

Sources: U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence;" Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*; Organization for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence; Defense Innovation Board, *Supporting Document*.[192]

Appendix II: Comparing U.S., French, and Australian lexicons for safe, ethical, and controlled AI in defense

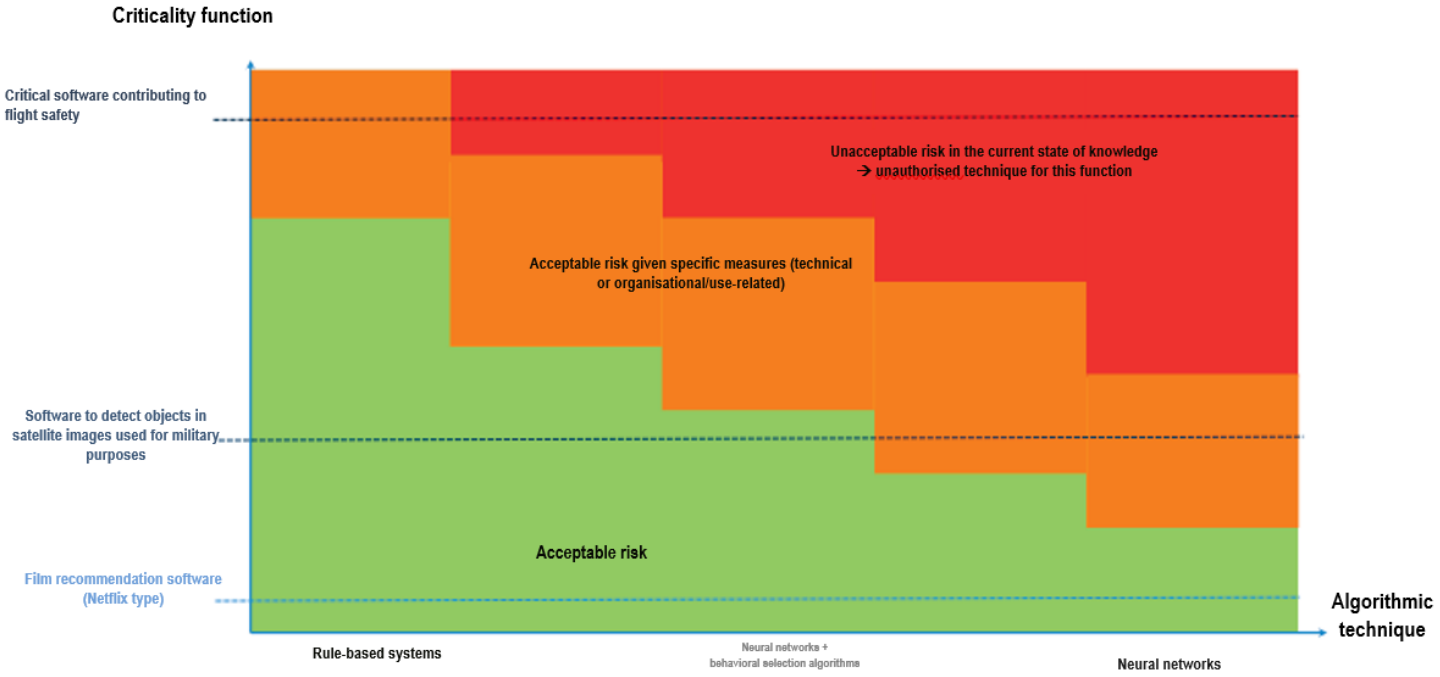
U.S. principles for safe and ethical AI *(adopted)*	French guidelines for controlled AI* *(from AI strategy)*	Australian facets of ethical AI in defense *(workshop result)*
"**Responsible.** DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities."	"**Assurance of trustworthy, controlled, and responsible AI.**" Aspects on responsibility are more implicit, but still have a functional approach to managing risk: "ensure that a 'right level' of trustworthiness and robustness is assessed for each AI application [...starting from] the design phase [...] to identify which of the different functions are the most critical in order to deduce the relevant requirements in terms of development, qualification and monitoring in use."	**Responsibility.** Who is responsible for AI? (Related topics: education, command.)
"**Equitable.** The Department will take deliberate steps to minimize unintended bias in AI capabilities."	Bias-derived risks mentioned as part of introduction to why AI needs to be more robust; only implicit in "a robust ethical and legal framework" section. (See "Reliable.")	Bias mitigation part of 'governance' facet & related topics of human factors, supply chain, and confidence. (See "Reliable.")

"**Traceable.** The Department's AI capabilities will be developed and deployed such that relevant personnel possess an appropriate understanding of the technology, development processes, and operational methods applicable to AI capabilities, including with transparent and auditable methodologies, data sources, and design procedure and documentation."	Data "configuration" (*i.e.*, documentation): "Preserving digital sovereignty therefore also involves controlling the algorithms and their configuration, and the governance of data," as well as "their learning elements, their combinations and their data." (Explainability touched on, but not main principle.)	**Traceability.** How are the actions of AI recorded? (Related topics: explainability and accountability)
"**Reliable.** The Department's AI capabilities will have explicit, well-defined uses, and the safety, security, and effectiveness of such capabilities will be subject to testing and assurance within those defined uses across their entire life-cycles."	"**Assurance of trustworthy, controlled, and responsible AI.**" Aspects on trust ("rigorous systems design which must guarantee total compliance with the human-defined framework, and on the ministry's capacity to evaluate and certify such systems.") Linked to robustness and resilience.	**Trust.** How can AI be trusted? (Related topics: sovereign capability, safety, supply chain, test & evaluation, misuse and risks, authority pathway, and data subjects.)
"**Governable.** The Department will design and engineer AI capabilities to fulfill their intended functions while possessing the ability to detect and avoid unintended consequences, and the ability to disengage or deactivate deployed systems that demonstrate unintended behavior."	**Governability of data.** Largely adoption-focused: abiding by security and ethical rules to maintain or even reinforce confidence by ministry personnel/armed forces and acceptability by the public; **Governance.** Relevant structures, including CCIAD, ministerial ethics committee, *et al.* Also education: "take steps to raise awareness among ministry staff of the use of AI, especially from an ethical standpoint."	**Governance.** How is AI controlled? (Related topics: effectiveness, integration, transparency, human factors, scope, confidence, and resilience.)
Ethical principles embedded in/building on legal framework, but not own principle.	Highlighted in framing, but not as principle (similar to U.S.).	**Law.** How can AI be used lawfully? (Related topics: protected

		symbols and surrender, and de-escalation.)
Source: U.S. Department of Defense	*Source: AI Task Force, French Ministry of Armed Forces*	*Source: Devitt et al., Australian Department of Defence*

* "Controlled" AI is the official translation of "maîtrisée" in the French document, which can broadly be seen as the need to harness AI and overcome its challenges. This is a separate term from "human control" and should not be conflated with topics related to command and control.

Appendix III: French risk level of AI-based algorithmic technologies according to criticality



Source: AI Task Force, French Ministry of Armed Forces

Appendix IV: Australian Method for Ethical AI in Defence – Risk Matrix

Risk Checklist			Risk Matrix	
Topic	**Description/Reference**		**Activities**	*Define the activity you are undertaking*
Describe the military context in which the AI will be employed	*Combat/Warfighting*: force application, protection, and sustainment; situational understanding *Enterprise-level and rear echelon functions*: personnel, enterprise logistics, business process improvement (see Appendix E of Method)		**Ethical Issues to be addressed**	*Indicate the ethical issue the activity is intended to address*
Explain the types of decisions supported by the AI	1903 Defence Act, and 'Critical Decision Analysis' 2013; and Taxonomy of decision problems (single once-off vs. sequential decision making; multi-decision-maker cooperative vs. non-cooperative/iterated vs. non-iterated/zero-sum vs. non-zero sum; consensus decision making; consensus decisions/social choice (see Appendix F of method)		**Risks**	*Estimate the risk to the project objectives if issue is not addressed*
Explain how the AI integrates with human operators to ensure effectiveness and ethical decision making in the anticipated context of use and	See topics particularly under the Governance and Trusted sections: Governance: effectiveness, integration, transparency, human factors, scope, confidence and resilience Trust: Sovereign capability, safety, supply		**Actions**	*Define specific actions you will undertake to support the activity*

countermeasures to protect against potential misuse	chain, test and evaluation, misuse and risks, authority pathway and data subjects			
Explain framework/s to be used	Examples given include: facets and topics from Method; Australian government's AI Ethics Principles; IEEE Ethically Aligned Design; US DOD Principles; etc.		**Timeline**	*Provide a timeline of the activity*
Employ subject matter experts to guide AI development	For example, use consultants, contractors, or hire employees with relevant expertise in military ethics, decision science, law, human factors, and data science to assist with AI project conceptualization and planning		**Outcome**	*Define action and activity outcomes*
Employ appropriate verification and validation techniques to reduce risk	Seek out best practice in autonomy and intelligent system test and evaluation methods to accelerate certification and assurance for acquisition, adoption and social license		**Assignee**	*Identify the responsible party/ies*
			Status	*Provide a status update*

Source: Synthesized from Devitt et al., Australian Department of Defence, 33, 57–60.

Appendix V: U.K. Dstl factors for success: consent and confidence

Topic	Guiding questions
Consent: (a) legal and regulatory constraints; (b) as well as policy, ethics, and risk appetite, (c) willingness of suppliers and partners to support where required	
Legal	(1) Are there any externally imposed constraints on our capability, such as legal and regulatory frameworks that we need to follow? (2) Have we checked the international position as well as domestic? (3) What do we need to stay

	within these constraints? (4) Is the legal position clear or ambiguous? Do we need to get advice to ensure we comply? (5) Is it possible to influence those constraints if we can't operate within them? (6) Note that anything involving legal matters will take longer than you can possibly imagine, so factor this in.
Policy and risk appetite	(1) Is the enterprise (including partners, suppliers and collaborators) likely to be willing to pursue this capability, based on its own internal policies and risk appetite? (2) What are the existing policy and risk positions of our organisations? (3) Are there international policies to consider? (4) What do we need to do to stay within these constraints? (4) Is the policy position clear or ambiguous? Do we need to get advice to ensure we comply? (5) Is it possible to influence the policy if we can't operate within it? (6) Should we try to influence this? For example, what are the risks of not developing the capability?
Ethics	(1) Fundamentally, should we pursue this capability? (2) Have we considered the ethics of doing so, and equally the ethics of not? (3) What is our organisation's existing ethical position? (4) Does this capability operate within that position? (5) Do our ethics align with those of our partners, and will these partners support and engage in our work? (6) Are systems fair and equitable?
Confidence: (a) satisfying regulatory and safety requirements; (b) inspiring trust through assurance, explainability and effective exercising; (c) being aware of the risks through an understanding of threats, vulnerabilities, means of failure and wider resilience	
Assurance	(1) Will we be able to certify that the system satisfies all relevant regulations, including safety and security standards? (2) Will all of the functions that the system performs work reliably, as expected and for as long as they need to? The latter is an important point if you have a learning system where the performance could change over time – how do you understand and maintain performance? (3) Do we have an understanding of behaviours the system must not have (*e.g.* harming people – this is generally considered to be a bad thing) and how they can be prevented? (4) Do we understand what level of assurance is required?
Trust	(1) Who needs to trust the system, what do they need to understand and what do you need to provide to obtain this trust? (2) This sounds like a simple question but can have many facets – there will be different trust considerations for the direct users, those making decisions based on its outputs, the regulators and the general public

Explainability	(1) Do we need to be able to explain why the AI made a particular decision; both at the time, and in retrospect? If so, how can we do this? This is another question that may impact on your algorithm selection: if you really need to know why the system produced a certain output, some types of algorithm will be more suitable than others.
Resilience	(1) Do we understand the vulnerabilities in the system, and the risks it might introduce to our operations or business? Will the system fail gracefully if it encounters situations beyond its design parameters?
Experimentation	(1) How suited is the system for experimentation, to build experience and confidence before it is used in a live environment?

Source: Adapted from: U.K. Defence Science and Technology Laboratory, 22–5.

Appendix VI: Canadian Military Ethics Assessment Framework for Human Enhancement Technologies

Compliance with National Laws and Codes of Conduct	Questions raised about whether a technology interferes with the common values, laws and expected behaviors that guide both military employees in all activities related to their professional duties
Jus Ad Bellum Principles	Questions raised about whether a technology disrupts Jus ad Bellum principles: criteria to be met before entering a conflict to ensure that all conflicts entered into are justified
Law of Armed Conflict/Jus in Bello Principles	Questions raised about whether a technology violates the international laws that must be followed during times of conflict to protect those affected by conflict and to regulate means of warfare
Health and Safety	Questions raised about direct or indirect impacts the enhancement may have on the physical or psychological well-being of a soldier or civilian
Accountability and Liability	Questions raised about risk and responsibility for enhancement failures or unanticipated and undesired effects of an enhancement
Privacy, Confidentiality, and Security	Questions raised about sharing, storing, and using information obtained by an enhancement, and security risks of an enhancement resulting from adversary detection or hacking
Equality	Questions raised about the influence of an enhancement on fairness and functionality within the CAF, between militaries and in society
Consent	Questions raised about whether the enhancement is mandatory or voluntary

Humanity	Questions raised about the influence of an enhancement on a soldier's morals and personhood
Reliability and Trust	Questions raised about how close the enhancement technology is to commercialization and use by the military, and remaining modifications required for usability on the battlefield
Effect on Society	Questions raised about how an enhancement will impact civilians and perception outside of the forces
Preparedness for Adversaries	Questions raised about how adversaries will view our use of enhancements and how adversaries may use enhancements themselves

Source: Girling et al., Defence Research Development Canada, 15.

Appendix VII: NATO STO technical report on stakeholders and the application of dimensions of human control in the use of force

Stakeholder groups that contribute to ensuring appropriate human control in fielded systems

Stakeholder Groups	Role in Delivering Appropriate Human Control in Future Systems [Effective Human Control = EHC; Meaningful Human Control = MHC]
Policy makers	Set policy, standards and doctrine to facilitate EHC and MHC across NATO systems
R&D/ Scientific Community	Identify and fill knowledge gaps in the field of EHC and MHC; Conduct research to provide evidence to support the specification and design of systems that support MHC and EHC
System Acquirers	Specify, contract and accept AI enabled systems that support EHC and MHC

System Designers	Conduct and analysis, use Human Centred Design approaches such as those described in ISO 9241 210 [0], and apply best practice to systems design and testing to develop systems that support the delivery of EHC and MHC
Organisational Users	Integrate AI enabled systems within their wider system of systems and organisational structures such that they enable EHC and MHC
End users	Train, support and employ human machine teams that deliver EHC and MHC; Deliver operating procedures and practices to deliver EHC and MHC

Application proposed dimensions of (ET) developed Dimension of human control to iPRAW's Requirements for Human Control in the Use of Force

	Situational Understanding	Intervention
Control by Design (Technical Control)	Design of systems that allows human commanders the ability to monitor information about environment and system	Design of systems with modes of operation that allow human intervention and require their input in specific steps of the targeting cycle based on their situational understanding
	The system design allows the Human to develop sufficiently accurate situation, and system awareness/understanding to identify risks to violating IHL and/or unacceptable moral, ethical or operational outcomes The system design allows the Human to predict the behavior of the system and its effects on the environment (physical and information)	The system design allows the Human to impact on the behavior of the system in time to prevent an undesirable act (violating IHL and/or unacceptable moral, ethical or operational outcomes)
Control in Use (Operational Control)	Appropriate monitoring of the system and the operational environment	Authority and accountability of human operators, teammates and commanders; abide by IHL
	The Human has sufficiently accurate situation, and system awareness/understanding to identify risks to violating IHL and/or unacceptable moral, ethical or operational outcomes.	The Human is able to exercise freedom of choice and has the ability to affect system behavior during use to ensure that accountability and

	The Human is able to predict the behavior of the system and its effects on the environment (physical and information)	adherence to IHL are maintained Training with systems allows users to understand and predict system behaviors across different situations in order to avoid undesirable outcomes or failure to comply with IHL. Organizational culture does not indirectly impact on freedom of choice and willingness to question system behaviors and actions

Source: Boardman and Butcher, NATO Science & Technology Organization, 9–11.

Appendix VIII: Applicability of civilian EU trustworthy AI principles to the defense sector

Ethical principle	Description	Convergence with defense sector	Divergence with defense sector
Respect for human autonomy	Ability for humans to maintain self-determination and partake in democratic process	Design of AI systems to augment, complement and empower human skills (alignment; human-machine teaming)	"Should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans" ≠ military justification to coerce etc.
Human agency and oversight	Assessment of fundamental rights impact prior to development Define human agency and governance with human at centre to prevent adverse effects	Similar to Article 36 review for autonomous systems as well as appropriate level of judgment/control, as well as TEVV for new military capabilities Consideration of legal aspects throughout development and deployment pipeline	N/A when applied to existing legal framework for armed conflict "Human-in-command" ("capability to oversee overall activity of the AI system" may conflict with select niche AI supervisor systems for non-combat effects requiring machine speed (*e.g.*, cyber etc.)

		Oversight with human in/on loop	
Prevention of harm	No exacerbation of harm or adverse effects on human beings	Ensure technically robust systems are not vulnerable to malicious use	Harm inherent to conduct of warfare; "AI systems and the environments in which they operate must be safe and secure" ≠ operational realities
Technical robustness and safety	Preventative approach to risks in AI development, including protection against adversarial attacks and mitigation of unintended harm	Preventing "unacceptable" harm consistent with militaries seeking to minimise "unintended" consequences Principles recommend promoting "convergence between the AI community and the security community"	Ensure AI system performs as intended "without harming living beings or the environment" ≠ conduct of warfare
Privacy and data governance	Data governance covering quality and integrity of data, access protocols and capabilities for processing that protect privacy	Mutual interests in assuring high level of data quality, integrity, and access Potential overlap in technical measures to protect privacy & protect confidential information	Potentially more divergence for implementing privacy and data governance, although already have carveouts for public safety, national security, and defense in GDPR
Societal and environmental well-being	Measurement and assessment of environmental/ecological impact, and societal impact including on democratic institutions	Measuring impact of AI adoption on "social agency" important for understanding role of human in human-machine teams, including impact on health and well-being of personnel; focus on "encouragement" rather than anything prescriptive	Social well-being not taken into account for adversaries (linking back to main tension between defense sector and prevention of intentional & unintentional harm)

Fairness	Just distribution of costs/benefits Freedom from bias, discrimination, and stigmatisation Accountability and ability to contest AI decisions	Cost-benefit analysis remains relevant Accountability and explicability of decision-making processes	"Should never lead to people being deceived" ≠ military aim of establishing "unfair" advantage (*e.g.*, operational surprise)
Diversity, non-discrimination and fairness	Avoid unfair bias by enabling inclusion and diversity Improve accessibility through user-centric, universal design Solicit stakeholder feedback throughout lifecycle	Place oversight processes to analyse and address system's purpose, constraints, requirements, and decisions to reduce bias in programming (and training phase) Use universal design to enable equitable access	"Unfair" not in interest of militaries who seek to establish "unfair advantage"; nonetheless important to consider bias especially for internal management systems or unintended harm
Accountability	Mechanisms to ensure responsibility, accountability, auditability for AI systems and their outcomes Redress for adverse impacts to ensure trust, including through documentation	Enable assessment of algorithms, data, and design processes and conduct independent audit of safety-critical applications Identify, assess, document, and minimise potential negative impacts of AI systems, including through red-teaming exercises prior to/during development	N/A [But noteworthy that documentation-centric definition of accountability differs from CCW 'human accountability' term]
Societal and environmental well-being	(See description under prevention of harm; defined as relevant to both principles)		
Explicability	Transparency and openness about processes and capabilities/purpose of AI system	Augmentation of other explicability measures (traceability, auditability, communication) for black-box AI	Limitations on transparency (which may affect feedback loops to ensure that humans can contest AI-made or AI-augmented decisions)

	Explainability of decisions to those affected		
Transparency	Ensure that data, systems, and business models are traceable, explainable, and communicated	Facilitate auditability and explainability through documentation to maximise reproducibility and minimise repetition of mistakes Have ability to explain trade-offs between explainability and accuracy	Inform humans when "they are interacting with an AI system" ≠ psychological dimension of strategy

Adapted from author's conference presentation (November 11, 2019); see also High-Level Expert Group on Artificial Intelligence, European Commission.

Appendix IX: Five Eyes TTCP Cyber Strategic Challenge Working Group attributes of trustworthiness of cyber systems

Main attributes of trustworthiness	Equivalent or sub-attribute(s) to the main attribute	Trust	Resilience	Agility
Reliability	An attribute of dependability; predictability; competence; consistency; stability; certainty; fault-forecasting; high-confidence; assurance; survivability	✓	✓	✓
Availability	An attribute of dependability; an attribute of security	✓	✓	✓
Safety	An attribute of dependability	✓	✓	
Confidentiality	An attribute of security	✓		

Integrity	An attribute of security; accuracy; credibility	✓		
Robustness	Fault-tolerance; performability; accountability; authenticity; nonrepudiability	✓	✓	✓
Maintainability	Recoverability; retainability; correctability; self-healing; self-repair		✓	
Adaptability	Autonomy; learning; extensibility; reconfigurability		✓	✓
Usability	Automatability; flexibility; learnability; satisfaction; compatibility; reusability; complexity			✓
Timeliness	Quickness; decisiveness			✓
Leanness	Efficiency; simplicity; scalability			✓
Reactiveness	Readiness; fault-removal		✓	✓
Proactiveness	Preparedness; fault-prevention			✓

Source: Replicated from: Cho et al., MILCOM 2016.

Ontology-based Trust, Resilience, and Agility Metrics (TRAM) Framework

Source: Replicated from: Cho et al., MILCOM 2016.

Endnotes

[1] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense* (Washington, DC: May 26, 2021), https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF.

[2] U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020, https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

[3] North Atlantic Treaty Organization, *North Atlantic Treaty*, April 4, 1949, https://www.nato.int/cps/en/natohq/official_texts_17120.htm and Katie Lange, "National Defense Strategy: Alliances and Partnerships," U.S. Department of Defense, October 8, 2018, https://www.defense.gov/Explore/Features/story/Article/1656016/national-defense-strategy-alliances-and-partnerships/.

[4] Denis Mercier, "SACT's opening remarks to the NAC/MC Away Day," Allied Command Transformation, March 22, 2018.

[5] Joanna van der Merwe, "NATO Leadership on Ethical AI is Key to Future Interoperability," *Center for European Policy Analysis*, February 17, 2021, http://cepa.org/nato-leadership-on-ethical-ai-is-key-to-future-interoperability/.

[6] Yuna Huh Wong, John M. Yurchak et al., *Deterrence in the Age of Thinking Machines* (Santa Monica, CA: RAND Corporation, 2020), 6, 60.

[7] This point has also been made in relation to the difficulties of a prematurely prohibitive ban on lethal autonomous weapon systems. Although autonomy is beyond the scope of the study, NSCAI Executive Director Yll Bajraktari makes a similar point on the relationship between ethics and interoperability: "The effects of a prohibition agreement likely would run counter to the U.S. strategic interests as commitments from states such as Russia and China are likely to be empty ones. So, the primary impact of an agreement would be to increase pressure on those countries that abide by international law, including the United States and its democratic allies and partners. If U.S. allies joined an agreement while the United States did not, the diversion would likely hinder allied military interoperability. That would be something really difficult for us and our allies. For these reasons, we believe that practical and strategic problems with a prohibition treaty outweigh the potential benefits for the United States and its allies and partners." Craig Smith and Yll Bajraktari, "Episode #071: AI and

National Security: US vs China," *Eye on AI*, May 5, 2021, transcript available at: https://www.eye-on.ai/podcast-071.

[8] Audley Genus and Andy Stirling, "Collingridge and the dilemma of control: Towards responsible and accountable innovation," *Research Policy* 47, no. 1 (February 2018): 61–69.

[9] It should be noted that this is different from the legal concept of state responsibility. In international law, state responsibility refers to principles that guide how a state is held accountable *after* a violation. As Appendix I shows, the DIB considered the doctrine of state responsibility (as "remediation mechanisms for actions after hostilities have ended") as a layer of responsible AI, albeit one external to DOD. Nevertheless, the concept is sufficiently different that it should not be confused with the more general concept of responsible state behavior as described in this report. See: Defense Innovation Board, *Supporting Document: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, DC: October 30, 2019), 28–29.

[10] While this may seem obvious, not all design begins with identifying a need before matching it with the appropriate solution—technological or otherwise. One example of ways to improve the role of the human is reducing her cognitive load and affording her more capacity to make accurate decisions that enhance compliance with international humanitarian legal principles.

[11] This runs parallel to determinations of human in, on, or out of the loop that are often discussed in relation to autonomy in weapons.

[12] The DIB describes the ethical and technical distinctions between AI and autonomy in its supporting document on AI ethics recommendations: "While some autonomous systems may utilize AI in their software architectures, this is not always the case. The interaction between AI and autonomy, even if it is not a weapon system, carries with it ethical considerations. Indeed, it is likely that most of these types of systems will have nothing to do with the application of lethal force, but will be used for maintenance and supply, battlefield medical aid and assistance, logistics, intelligence, surveillance and reconnaissance, and humanitarian and disaster relief operations. Various ethical dimensions may arise depending upon the system and its domain of use, and those will change depending upon context." See: Defense Innovation Board, *Supporting Document: AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, 10.

[13] Several of these types of systems could be integrated into a weapon system with a degree of autonomy. Still, the differences between the autonomy-related and AI-related risks in that system would still apply.

[14] For a comprehensive summary and links to country positions, see: Dustin E. Lewis (editor), "A Compilation of Materials Apparently Reflective of States'

Views on International Legal Issues Pertaining to the Use of Algorithmic and Data-Reliant Socio-Technical Systems in Armed Conflict," *Harvard Law School Program on International Law and Armed Conflict*, December 2020, https://pilac.law.harvard.edu/a-compilation-of-materials-apparently-reflective-of-states-views-on-international-legal-issues-pertaining-to-the-use-of-algorithmic-and-data-reliant-socio-technical-systems-in-armed-conflict.

[15] Initial discussions began in 2013, and subsequently a Group of Governmental Experts was established to focus on LAWS from 2016–2021.

[16] While the governments have not issued approaches, there is pressure in both countries to boycott military research. The Korea Advanced Institute for Science and Technology (KAIST), for instance, became the subject of controversy when researchers called for a boycott because of KAIST's partnership with the defense contractor Hanwha, which is involved in building autonomous weapons. The following year, KAIST incorporated an AI ethics lecture into its computer ethics course—but it does not touch on questions related to defense, security, or public order. Likewise, the Science Council of Japan has also encouraged Japanese universities and research institutions to set up ethical screening mechanisms that block military research, though not specific to AI. This builds on a historical, post-war movement, and in the first year, 46 of 183 institutions followed through on establishing the ethical screen. Still, the issues give a sense of domestic limitations that ideally should motivate more established government policies addressing public concerns. Separately, it is worth noting that both South Korea and Japan are part of the JAIC-led PfD that is also covered in this report. See: Shin Yoo, "AI and Ethics," Korea Advanced Institute for Science and Technology, https://coinse.kaist.ac.kr/assets/files/teaching/cs489/cs489-slide07.pdf. James Vincent, "Leading AI researchers threaten Korean university with boycott over its work on 'killer robots,'" *The Verge*, April 4, 2018. https://www.theverge.com/2018/4/4/17196818/ai-boycot-killer-robots-kaist-university-hanwha; Yojana Sharma, "Urgent need to address ethics of artificial intelligence," *University World News*, July 3, 2018, https://www.universityworldnews.com/post.php?story=20180703090452302; "1 year on, Japan science council's rejection of military research has little traction," *The Mainichi*, March 30, 2018, https://mainichi.jp/english/articles/20180330/p2a/00m/0na/010000c; and Suvendrini Kakuchi and Aimee Chung, "Military use of research pushback in Japan, South Korea," *University World News*, April 5, 2018, https://www.universityworldnews.com/post.php?story=20180405155744230.

[17] Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense* (Washington, DC: October 30, 2019), 2. https://media.defense.gov/2019/Oct/31/2002204458/-1/-1/0/DIB_AI_PRINCIPLES_PRIMARY_DOCUMENT.PDF.

[18] Defense Innovation Board, *AI Principles*, 4; Defense Innovation Board, *Supporting Document*, 2019, 48–52.

[19] Defense Innovation Board, *AI Principles,* 4.

[20] Defense Innovation Board, *AI Principles,* 4; Defense Innovation Board, *Supporting Document*, 21–22.

[21] Defense Innovation Board, *AI Principles,* 4–6.

[22] There are slight differences in the language, including: removing the term "harm" (*e.g.* unintended harm, inadvertent harm) for equitable and governable AI; ensuring "relevant personnel" and not just technical experts understand and can trace the technology; and choosing to not specify if disengagement and deactivation of systems is "human or automated." See: Defense Innovation Board, *AI Principles*, 4–6; U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," 2020.

[23] U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence."

[24] Dana Deasy and Jack Shanahan, "Transcript: Department Of Defense Press Briefing on the Adoption of Ethical Principles for Artificial Intelligence," U.S. Department of Defense, February 24, 2020, https://www.defense.gov/Newsroom/Transcripts/Transcript/Article/2094162/department-of-defense-press-briefing-on-the-adoption-of-ethical-principles-for/.

[25] The initial cohort involved 15 people from across the AI lifecycle, including personnel involved in "product design and development, testing and evaluation/verification and validation, and acquisition teams as well as policy, plans, and performance." See: Joint Artificial Intelligence Center, "Department of Defense Joint Artificial Intelligence Center Responsible AI Champions Pilot," August 21, 2020, https://www.ai.mil/docs/08_21_20_responsible_ai_champions_pilot.pdf.

[26] Brandi Vincent, "Pentagon Confirms Alka Patel to Lead the Implementation of Its New Ethical AI Principles," *Nextgov*, February 27, 2020, https://www.nextgov.com/emerging-tech/2020/02/pentagon-confirms-alka-patel-lead-implementation-its-new-ethical-ai-principles/163386/.

[27] Deasy and Shanahan, "Transcript."

[28] Jackson Barnett, "Why the Pentagon can't go it alone on AI," *FedScoop*, April 24, 2020, https://www.fedscoop.com/experts-urge-us-nato-not-to-go-it-alone-on-developing-artificial-intelligence/.

[29] Barnett, "Why the Pentagon can't go it alone on AI."

[30] See Appendix I for a comparison of what RAI means in the U.S. context.

[31] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense* (Washington, DC: May 26, 2021), 2, https://media.defense.gov/2021/May/27/2002730593/-1/-1/0/IMPLEMENTING-RESPONSIBLE-ARTIFICIAL-INTELLIGENCE-IN-THE-DEPARTMENT-OF-DEFENSE.PDF

[32] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*, 2.

[33] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*, 2.

[34] The complete list of RAI "foundational tenets" includes: RAI governance, warfighter trust, AI product and acquisition lifecycle, requirements validation, responsible AI ecosystem, and AI workforce. See: Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*, 2.

[35] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*, 3.

[36] For example, investments in microelectronics stem from supply chain concerns (not only for AI) that are separate from discussions on ethics.

[37] Defense Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapon Systems* (Paris: French Ministry of Armed Forces, April 29, 2021), 32.

[38] This stems from a stronger overall emphasis on AI in defense than other European allies, as also seen in the national AI strategy designating security and defense as one of its four priorities for industrial policy. See: Ulrike Franke, "Not smart enough: The poverty of European military thinking on artificial intelligence" (European Council on Foreign Relations, December 18, 2019), https://ecfr.eu/publication/not_smart_enough_poverty_european_military_thinking_artificial_intelligence/?s=08.

[39] AI Task Force, *Artificial Intelligence in Support of Defence* (Paris, France: French Ministry of Armed Forces), September 2019, 7–10, https://www.defense.gouv.fr/salle-de-presse/communiques/communique_publication-du-rapport-du-ministere-des-armees-sur-l-intelligence-artificielle.

[40] The French iteration of controlled AI has several components: maintaining freedom of action and interoperability with allies; assuring the ethical concepts of

trustworthiness and responsibility; ensuring the resilience and upgradeability of systems; designing systems that can be upgradeable over decades; and preserving digital sovereignty to maintain confidentiality and control over information. See: AI Task Force, *Artificial Intelligence in Support of Defence*, 9–10.

[41] This concept of harnessing and governing AI also extends to areas such as arms control. Human control is encompassed more in the French Ministry of Armed Forces conception of responsible AI.

[42] AI Task Force, *Artificial Intelligence in Support of Defence*, 9.

[43] The former NATO Legal Adviser has alternatively referred to French ethical AI principles for defense as "(1) respect for international law; (2) the presence [of] sufficient human control; and (3) ensuring responsibility of human command." These are certainly themes present in the document, but France does not refer to them as principles, and aspects are more related to autonomy in weapons than AI itself. The third of these is discussed in relation to responsible AI. See: Steven Hill, "AI's Impact on Multilateral Military Cooperation: Experience from NATO," *American Journal of International Law Unbound* 114 (April 27, 2020): 148.

[44] AI Task Force, *Artificial Intelligence in Support of Defence*, 7.

[45] In addition to Appendix III, a simpler view on this could also be designating safety-critical and mission-critical systems—which are already subject to different engineering standards for weapon systems.

[46] AI Task Force, *Artificial Intelligence in Support of Defence*, 11.

[47] The strategy sees bias as a risk in at least two ways: unrepresentative learning data making systems less robust, and what it terms "voluntary bias," whereby third parties attack training data and models, including possibly on request. AI Task Force, *Artificial Intelligence in Support of Defence*, 7.

[48] AI Task Force, *Artificial Intelligence in Support of Defence*, 5, 9.

[49] More broadly, the French national AI strategy—which included a French defense procurement engineer on the six-person drafting council—issued a basis of five principles for its approach to ethical AI: (1) transparency and auditability concerning autonomous systems, (2) protection of rights and freedom, (3) responsibility and accountability, (4) diversity and inclusion, and (5) purposeful politicization of the role of the technology in society. See: Cédric Villani, *Towards a French and European Strategy* (Paris: French Government, 2018), 113–114, https://www.aiforhumanity.fr/pdfs/MissionVillani_Report_ENG-VF.pdf.

[50] The document suggests standardization and incorporating these questions into the front end of software engineering (as a form of "ethical design") as the answer, but does little to acknowledge the challenges beyond saying that "Merely verifying performance is not enough." Villani, *Towards a French and European Strategy*, 113–114.

[51] The Ministry's "Digital Ambition" is the primary document on data governance, but similarly does not specify data ethics apart from a few references to trust and protection of personal data. See: French Ministry of Armed Forces, *Ambition Numérique du Ministère des Armées* (Paris, France: 2019), 22–3; Villani, *Towards a French and European Strategy*, 13.

[52] French Ministry of Armed Forces, *Ambition Numérique du Ministère des Armées*, 10.

[53] The focus here is more on preventing and correcting undesirable outcomes, rather than reproducing positive ones. French Ministry of Armed Forces, *Ambition Numérique du Ministère des Armées*, 13.

[54] The French imperative to maintain technological independence is stronger than any other European ally, and largely motivates French defence industrial policy as well as the political agenda of "strategic autonomy" and "digital sovereignty" at the national and European levels. Other documents that reinforce this include those referenced in footnotes 47 and 49, as well as the 2019 Defence Innovation Orientation Document (2019) and the Ministry of Armed Forces Digital Transformation: Key Concepts (2020).

[55] AI Task Force, *Artificial Intelligence in Support of Defence*, 10.

[56] This strong language intends to set the political tone for adoption, and is not purely about ethics. Further, while the "stranglehold" motivates sovereignty, there are few specifics in the strategy about hardware components or cloud capabilities, beyond the recognition that these are not French or European strengths. Auditability is only tied, here, to models and data. AI Task Force, *Artificial Intelligence in Support of Defence*, 10, 24.

[57] AI Task Force, *Artificial Intelligence in Support of Defence*, 10.

[58] It is also noteworthy for readers interested in cooperation that freedom of action includes maintaining interoperability with allies. See: AI Task Force, *Artificial Intelligence in Support of Defence*, 9, 14.

[59] The French strategy notes that law and ethics are "incorporated into the strict and sequenced process of planning the use of force and into a chain of decision-making for the application of force established by the rules of engagement,

validated by government." See: AI Task Force, *Artificial Intelligence in Support of Defence*, 10.

[60] The French military AI strategy does comment on autonomous weapons, mainly to reiterate the formal position that France will not develop "fully autonomous systems where human operators have no control over the definition and performance of their missions." See: AI Task Force, *Artificial Intelligence in Support of Defence*, 10.

[61] AI Task Force, *Artificial Intelligence in Support of Defence*, 10.

[62] The fact that other uses of AI are beyond the scope of what the French Defence Ethics Committee has thus far studied is a substantively different starting point than the DIB considered in their 15-month process. One possible explanation is that the French Ministry of Armed Forces had already introduced their approach in the AI Strategy itself, and saw them as less urgent relative to hot-button issues of autonomy in weapons and the less-discussed questions of human augmentation. Relatedly, another explanation could be that France would prefer to dedicate its capital to push AI ethics for defense at the multilateral (EU) level. See: Defence Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapon Systems* (Paris: French Ministry of Armed Forces, April 29, 2021), 3, 16.

[63] Defence Ethics Committee, "Avis portant sur le soldat augmenté" (Paris: French Ministry of Armed Forces, September 18, 2020), 6, https://www.defense.gouv.fr/salle-de-presse/communiques/communique_le-comite-d-ethique-de-la-defense-publie-son-avis-sur-le-soldat-augmente.

[64] On higher degrees of autonomy in decision-making, the Committee comments on adverse alterations of human judgment such as the human "feeling less involved in open fire procedures, causing detachment and loss of humanity in combat actions." See: Defence Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapon Systems*, 25.

[65] Xavier Dubreuil, "Quelle éthique des nouvelles technologies dans la guerre?" (French Command Doctrine and Education Center, June 26, 2020), https://www.penseemiliterre.fr/plugins/cdec/pdf/to_pdf.php?entry=379 and Florian Morilhat, *Éthique et Puissance Aérienne* (Paris, France: Economica, 2020); Alexandre Jubelin, "Donner la mort depuis les airs," interview with Florian Morilhat, *Le Collimateur*, March 2, 2021, https://www.irsem.fr/le-collimateur/donner-la-mort-depuis-les-airs-02-03-2021.html.

[66] This definition concerns soldiers, but opens up questions about other domains as well. For example, a French helicopter pilot has argued that pilots bear a heavier ethical burden as one or two operators in the air in comparison to the

collectivity of forces on ground or at sea. See (in French): Morilhat, *Éthique et Puissance Aérienne*.

[67] Translation from: Chief of Army, *L'exercise du métier des armes dans l'armée de terre: Fondements et principes* (Paris: French Army, January 1999), 4, https://defense.ac-versailles.fr/IMG/pdf/pistes_ethique_fondements.pdf.

[68] One French Army colonel has written (in French) that "only death or suffering can bend political will. At all costs, the soldier must remain at the heart of conflict and remain in charge of fires, supported by his ethics and moral strength which allow him to make the best decision, at the risk of his own life if necessary. The loss of tactical effectiveness will thus be compensated by the preservation of the moral, and therefore political, legitimacy of military action." Translated from: Chief of Army, *L'exercise du métier des armes dans l'armée de terre*, 4–5.

[69] Dubreuil, "Quelle éthique des nouvelles technologies dans la guerre?"

[70] Dubreuil, "Quelle éthique des nouvelles technologies dans la guerre?"

[71] Kate Devitt and Michael Gan et al., *A Method for Ethical AI in Defence* (Canberra: Australian Department of Defence, Defence Science and Technology Group, 2020), iii.

[72] Wong et al., *Deterrence in the Age of Thinking Machines*, 6, 60.

[73] Wong et al., *Deterrence in the Age of Thinking Machines*, 32.

[74] Devitt et al., *A Method for Ethical AI in Defence*, iii.

[75] The Data Ethics Canvas is not explicitly mentioned in the Method document, but features in other educational information about the Australian approach. See: Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Animation Summary," YouTube, June 6, 2021, https://www.youtube.com/watch?v=D_kwQZUth88; Devitt et al., *A Method for Ethical AI in Defence*, iii.

[76] Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Animation Summary"; Devitt et al., *A Method for Ethical AI in Defence*, iii.

[77] Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed," YouTube, June 6, 2021, https://www.youtube.com/watch?v=I1Cu4_murdE; Devitt et al., *A Method for Ethical AI in Defence*, 20–21.

[78] An Article 36 review is an example of the type of planning the government would need to know about in advance. LEAPP is a "data item descriptor." The United States uses a variation of this term, "data item description," which is a "standardization document that defines the data [content, format, and intended use] required of a contractor." See: Defense Standardization Program, "Frequently Asked Questions (FAQs): Data Item Descriptions," U.S. Department of Defense, accessed July 30, 2021, https://www.dsp.dla.mil/Policy-Guidance/FAQs/Data-Item-Descriptions/; Devitt et al., *A Method for Ethical AI in Defence*, 34, 62.

[79] For an overview of the scope and function of Article 36 reviews, see: International Committee of the Red Cross, "A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977," *International Review of the Red Cross* 88, no. 864 (December 2006), 931–956, https://www.icrc.org/en/doc/assets/files/other/irrc_864_icrc_geneva.pdf.

[80] More specifically, as Dr. Lauren Sanders of the International Weapons Review describes, technical measures like the pixilation of feed, whether and how identification of combatants is possible, and if those identification parameters are adjustable for different settings, would help determine what types of functions (targeting, intelligence, decision support) would be acceptable. See the segment starting at 25:38 of: Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed."

[81] Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed."

[82] One reason that the Australian government centered on contestability as one of their eight AI ethics principles is the controversy around the "Robodebt" scheme in 2016, in which the government used a faulty debt-assessment algorithm. The program issued inaccurate debt notices to welfare recipients. The debts have required significant effort to repay and the program's lawfulness has been questioned. Equally relevant here, the effects on society have been harmful to the mental health of citizens. While this was an automated, not AI system, its adverse impact on the Australian population made clear the priority to include legal recourse for algorithmic decision-making. See: Jordan Hayne and Matthew Doran, "Government to pay back $721m as it scraps Robodebt for Centrelink welfare recipients," *Australia Broadcasting Corporation*, May 29, 2020, https://www.abc.net.au/news/2020-05-29/federal-government-refund-robodebt-scheme-repay-debts/12299410; Monique Mann, "Technological Politics of Automated Welfare Surveillance: Social (and Data) Justice through Critical Qualitative Inquiry," *Global Perspectives* 1, no. 1 (June 19, 2020), https://doi.org/10.1525/gp.2020.12991.

[83] Australian Department of Industry, Science, Energy and Resources, "AI Ethics Principles," Australian Government, accessed August 2021, https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles.

[84] Henrietta Lyons, Eduardo Velloso, and Tim Miller, "Conceptualising Contestability: Perspectives on Contesting Algorithmic Decisions," arXiv preprint arXiv:2103.01774 (2021), https://arxiv.org/pdf/2103.01774.pdf.

[85] To be clear, there are Court Martial and Defence Force Magistrate proceedings, and a strict legal framework is established in the Defence Force Discipline Act of 1982. Coalition and Labor governments have unsuccessfully attempted to establish an Australian Military Court, the merits of which are beyond the scope here. The main point here is that the accountability and liability structures are different to that of the United States. See: David Letts, "ADF ain't broke, don't fix it," *The Sydney Morning Herald*, November 29, 2013, https://www.smh.com.au/opinion/adf-aint-broke-dont-fix-it-20131128-2ycru.html.

[86] The mateship ethos is core to Australian leadership doctrine. Further, as Michael Evans of the Australian Defence College describes in his analysis of political culture, the Australian egalitarianist ethos of "mateship and 'fair go'" put more focus on egalitarianism, social harmony, and the common good, in contrast to the "ideal of individual self-expression as enshrined in the American bill of rights." This focus on mateship "tend[s] to reflect a strongly legalistic frame of mind" and differs from the "principles of military professionalism" with more focus on the common good and the tradition of "defence without militarism." See: Australian Defence Force, *ADF Philosophical Doctrine: 0 Series | Command: ADF Leadership* (Canberra: Commonwealth of Australia, 2021), ADF-P-0 ADF Edition 3, 3, https://theforge.defence.gov.au/sites/default/files/2021-06/adf-philosophical-doctrine-adf-leadership.pdf via https://theforge.defence.gov.au/adf-philosophical-doctrine-adf-leadership; Michael Evans, "Towards an Australian way of war: Culture, politics, and strategy, 1901-2004," *Australian Army Journal* 31, no. 1 (June 2014): 181, 189.

[87] Devitt et al., *A Method for Ethical AI in Defence*, 27.

[88] This MinAI concept, which deals with AI-enabled decision support, is contrasted against MaxAI, which has a much higher level of autonomy and therefore is not discussed here. For more, see: Jason Scholz and Jai Gaillott, "AI in Weapons: The Moral Imperative for Minimally-Just Autonomy," *Journal of Indo-Pacific Affairs* (2019). See also Arthur Holland Michel's work on the difference between lethal and lethality-enabling autonomous weapon systems: Arthur Holland Michel, "The Killer Algorithms Nobody's Talking About," *Foreign Policy*, January 20, 2020, https://foreignpolicy.com/2020/01/20/ai-autonomous-weapons-artificial-intelligence-the-killer-algorithms-nobodys-talking-about/.

[89] As described in the article, "Athena AI is an artificial intelligence system that identifies and classifies objects and locations on a battlefield and communicates to the soldier which ones must not be targeted for legal or humanitarian reasons. This encompasses people such as enemy troops who have surrendered and civilians, as well as locations such as hospitals or other protected sites." The article also notes that Athena AI worked with ethicists and moral philosophers to develop a legal and ethical framework for the system before beginning development. See: Jonathan Bradley, "Athena AI helps soldiers on the battlefield identify protected targets," *create*, April 26, 2021, https://createdigital.org.au/athena-ai-helps-soldiers-identify-protected-targets/.

[90] Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed."

[91] Trusted Autonomous Systems Defence Cooperative Research Centre, "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed."

[92] Devitt et al., *A Method for Ethical AI in Defence*, 50.

[93] The Australian investigation of war crime allegations in Afghanistan also forms part of the context around ethics and state responsibility, and is beyond the scope here except to say that the redress mechanisms are part of the legal approach to state responsibility. See also footnotes 10 and 81.

[94] Her Majesty's Government, *Global Britain in a competitive age: The Integrated Review of Security, Defence, Development and Foreign Policy* (London: March 2021), 39–40, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/975077/Global_Britain_in_a_Competitive_Age-_the_Integrated_Review_of_Security__Defence__Development_and_Foreign_Policy.pdf.

[95] That said, one discrepancy is that the Integrated Review also calls into question the pace of global governance in relation to technology rule, standard, and norm development. See: Secretary of State for Defence, *Defence in a competitive age* (London: Ministry of Defence, March 2021), 8., https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974661/CP411_-Defence_Command_Plan.pdf; Her Majesty's Government, *Global Britain in a competitive age*, 30; Lord Clement-Jones, "Government must resolve AI ethical issues in the Integrated Review," *The House*, April 30, 2021, https://www.politicshome.com/thehouse/article/government-must-resolve-ai-ethical-issues-in-the-integrated-review.

[96] Secretary of State for Defence, *Defence in a competitive age*, 61.

[97] Her Majesty's Government, *Global Britain in a competitive age*, 73.

[98] U.K. Department for Digital, Culture, Media and Sport, *Government Response to the House of Lords Select Committee on Artificial Intelligence* (London: February 2021), 13, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/963696/Government_Response_to_the_HoL_Select_Committee_on_AI_v2.pdf; Her Majesty's Government, *Global Britain in a competitive age*, 73; Lord Clement-Jones, "Government must resolve AI ethical issues in the Integrated Review."

[99] Gavin Pearson, Phil Jolley, and Geraint Evans, "A Systems Approach to Achieving the Benefits of Artificial Intelligence in UK Defence," arXiv preprint arXiv:1809.11089 (2018), https://arxiv.org/pdf/1809.11089.pdf; Fiona Butcher, "HFM-300 Symposium on Human Autonomy Teaming (HAT) Trust Me I'm Artificial Intelligence," NATO Science & Technology Organization, April 11, 2019; Kenneth Payne, "Artificial Intelligence: The Challenges Facing UK Defence," *King's College London*, May 24, 2021, https://www.kcl.ac.uk/artificial-intelligence-the-challenges-facing-uk-defence.

[100] Dstl is an executive agency of the Ministry of Defence with five divisions—one of which is cyber and information systems, where the AI capability area has been housed since 2018. See: Steven Meers, "Challenges around socio-technical AI Systems in Defence: A Practitioners Perspective," U.K. Defence Science and Technology Laboratory, July 6, 2020.

[101] Ryan Daws, "DSTL will run Ministry of Defence's AI research lab," *AI News*, May 29, 2018, https://artificialintelligence-news.com/2018/05/29/dstl-ministry-of-defence-ai-research-lab/.

[102] Beth Wood, "The Turing appoints Dstl Ethics Fellow Mariarosaria Taddeo," The Alan Turing Institute, May 13, 2020, https://www.turing.ac.uk/news/turing-appoints-dstl-ethics-fellow-mariarosaria-taddeo.

[103] Multi-stakeholder engagement on AI ethics for defense is ongoing, including in forums not led by Dstl. A separate conference on "The role of AI in defence" in May 2021 focused on questions including: "[1] Where could and possibly shouldn't AI and ML be applied? [2] What are the ethical implications of choosing not to use AI technology? [3] What is the role of AI in making safety critical engineering decisions? [4] Addressing potential bias in algorithmic decision-making-What is the potential impact and how do we reduce the risk of bias in our datasets?" See: "Ethics in AI (Defence) Conference 2021," TechUK, May 26, 2021, https://www.techuk.org/what-we-deliver/events/ethics-in-ai-defence-conference-2021.html and Defence Science and Technology Laboratory, "Hundreds of the world's top minds debate ethics at Dstl's AI Fest,"

U.K. Government, November 4, 2020, https://www.gov.uk/government/news/hundreds-of-the-worlds-top-minds-debate-ethics-at-dstls-ai-fest.

[104] U.K. Defence Science and Technology Laboratory, *Building Blocks for Artificial Intelligence and Autonomy: A Dstl Biscuit Book* (Salisbury, U.K.: 2020), 20–25, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/955917/20201013-Building_Blocks_for_Artificial_Intelligence_and_Autonomy_v1.0.pdf.

[105] On a related note, the U.K. Government Communications Headquarters (GCHQ) has also issued guidance on AI ethics in intelligence. They are developing an "AI Ethical Code of Practice" which may dovetail with efforts here. See: Government Communications Headquarters, *Pioneering a New National Security: The Ethics of Artificial Intelligence* (Cheltenham, U.K.: 2021), 31, https://www.gchq.gov.uk/artificial-intelligence/index.html.

[106] U.K. Ministry of Defence, *Digital Strategy for Defence Delivering the Digital Backbone and unleashing the power of Defence's data* (London: May 27, 2021), 38, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/990114/20210421_-_MOD_Digital_Strategy_-_Update_-_Final.pdf.

[107] It was initially designed for human enhancement. Kimberly Girling, Joelle Thorpe, and Alain Auger, "Identifying Ethical Issues of Human Enhancement Technologies in the Military" (Defence Research Development Canada, 2017), 16.

[108] Michael Karlin, "Responsible AI in a national defence context," *Medium*, August 15, 2019, https://medium.com/@supergovernance/responsible-ai-in-a-national-defence-context-4de9ed99e34d.

[109] Girling et al., *Identifying Ethical Issues of Human Enhancement Technologies in the Military*, 16.

[110] Girling et al., *Identifying Ethical Issues of Human Enhancement Technologies in the Military*, 15.

[111] Department of National Defence of Canada, *The Department of National Defence and Canadian Armed Forces Data Strategy* (Ottawa: December 3, 2019), 8, https://www.canada.ca/en/department-national-defence/corporate/reports-publications/data-strategy.html.

[112] Department of National Defence of Canada, *The Department of National Defence and Canadian Armed Forces Data Strategy*, 19.

[113] Tom Cardoso and Bill Curry, "National Defence skirted federal rules in using artificial intelligence, privacy commissioner says," *The Globe and Mail*, February 7, 2021, https://www.theglobeandmail.com/canada/article-national-defence-skirted-federal-rules-in-using-artificial/.

[114] Separately, Canada has also seen other controversies about military AI. Canadians learned through the media, rather than through government transparency, that the military was using facial recognition technology from the company Clearview AI—which then led to it being deemed illegal and banned. See: Wendy Gillis, Alex Boutilier, and Kate Allen, "MPs call for parliamentary investigation as Canadian military and police forces confirm they've tried facial-recognition technology," *Toronto Star*, February 28, 2020, https://www.thestar.com/politics/federal/2020/02/28/mps-call-for-parliamentary-investigation-as-military-and-police-forces-confirm-theyve-tried-facial-recognition-technology.html and Kashmir Hill, "Clearview AI's Facial Recognition App Called Illegal in Canada," *The New York Times*, February 3, 2021, https://www.nytimes.com/2021/02/03/technology/clearview-ai-illegal-canada.html.

[115] See Requirement 6.1 of the Canadian Directive on Automated Decision-Making at: https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

[116] This national AI strategy from before the pandemic mentions that it would be completed in 2020, but no statements affirm that this timeline was met. Government of the Netherlands, *Strategic Action Plan for Artificial Intelligence* (Amsterdam: October 9, 2019), 16, https://www.government.nl/documents/reports/2019/10/09/strategic-action-plan-for-artificial-intelligence.

[117] The author thanks Maaike Verbruggen for helping translate question 12 and its answer in the following document: Dutch Ministry of Defence, "Answering written questions on the meeting of NATO defense ministers on February 17 and 18, 2021," February 10, 2021, https://www.rijksoverheid.nl/documenten/kamerstukken/2021/02/10/beantwoording-kamervragen-over-bijeenkomst-navo-ministers-defensie-17-18-februari-2021.

[118] Roughly translated (using DeepL), the risks (some of which are more related to economic security and Dutch normative power) include: overdependence on foreign technology companies that could abuse data, large-scale financial fraud, public opinion, deepfakes/undue influence on the public, control being overridden, inaccurate decisions/error, enhancement of malicious actors' existing capabilities, system interference, AI "lock-in" (too dependent on the systems), loss of control, export/diffusion of surveillance technology, availability and

accessibility, and the position of the Netherlands in the world. See: R.M. Neef and C.E.A. van Weerd, *Artificial Intelligence in the Context of National Security - Final Note Study within the National Security Analyst Network* (The Hague: Netherlands Organisation for Applied Scientific Research, 2020), 29–33.

[119] The report uses different language to make this point. Namely, it looks at aspects like public opinion, undue influence on the public, and the position of the Netherlands in the world. Neef and van Weerd, *Artificial Intelligence in the Context of National Security*, 29–33.

[120] Other allies may have similar technical reports that inform how they think about safety and ethics of AI. But few other allies have publicly announced dedicated AI strategies in which risk assessments and governance structures dealing with ethics and safety can be taken into account.

[121] Opposition to LAWS is written into the 2018 coalition agreement—and the 2014 coalition agreement before it also touched on the necessity to approach ethics as part of procurement and export decisions. Depending on the makeup of the coalition after the 2021 election, it is possible that language on ethics and military AI may be written into the Coalition treaty based on these precedents.

[122] The author thanks Torben Schütz for this point.

[123] Franke, "Not smart enough"; Zhijiang Zhao, "Germany Needs to Consider Military AI," *American Institute for Contemporary German Studies*, September 23, 2020, https://www.aicgs.org/2020/09/germany-needs-to-consider-military-ai/.

[124] As Zhao notes, German sources have suggested that the "Ministry of Defence does not receive funds from the national AI budget allocations." See: Zhao, "Germany Needs to Consider Military AI."

[125] The author thanks Simona Soare for this point.

[126] The report represents the views of the two authoring agencies, each part of their ministries of defense, but are not official defense- or government-wide views. Pages 45–51 focus on ethical implications. See: Development, Concepts and Doctrine Centre, *Human Augmentation– The Dawn of a New Paradigm* (Shrivenham: U.K. Ministry of Defence, May 13, 2021), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/986301/Human_Augmentation_SIP_access2.pdf.

[127] Development, Concepts and Doctrine Centre, *Human Augmentation*, 3; Sophia Becker, Christian Mölling, and Torben Schütz, "Learning together: UK-Germany cooperation on military innovation and the future of warfare" (Hanns Seidel Foundation, the Policy Institute, and King's College London, 2021),

https://dgap.org/sites/default/files/article_pdfs/uk-germany_military_innovation_.pdf.

[128] This German multi-stakeholder approach to Ethical Design/Responsible Use of Artificial Intelligence in FCAS is distinct from the United States' DOD-driven and DOD-led approach. For example, the inclusion of theologians is indicative of both the desire to have a holistic approach and the reality that some cultural aspects are more foreign to the United States.

[129] FCAS Forum, "Ethics of defence: On the responsible use of new technologies in a Future Combat Air System," 2020, www.fcas-forum.eu/en/articles/ethik-und-verteidigung/.

[130] Florian Keisinger and Wolfgang Koch, "Defence and responsibility. How can we ensure that new technologies are used responsibly in a Future Combat Air System?," manuscript for *Behörden Spiegel*, prepared in April 2020, http://www.fcas-forum.eu/press/Op-ed-Keisinger-Koch-Behoerden-Spiegel-Defence-and-responsibility.pdf and https://www.behoerden-spiegel.de/2020/11/10/ethik-neuer-technologien-in-einem-future-combat-air-system/.

[131] Keisinger and Koch, "Defence and responsibility"; FCAS Forum, "Experts panel," accessed June 2021, www.fcas-forum.eu/en/experts.

[132] In addition to the cross-referencing of the trustworthy AI principles for defence in Appendix VIII, readers can see the ALTAI web tool at: European AI Alliance, "ALTAI – The Assessment List on Trustworthy Artificial Intelligence," European Commission, https://futurium.ec.europa.eu/en/european-ai-alliance/pages/altai-assessment-list-trustworthy-artificial-intelligence.

[133] Dirk Hoke, "The Responsible Use of Artificial Intelligence in the Future Combat Air System," LinkedIn, April 7, 2021, https://www.linkedin.com/pulse/responsible-use-artificial-intelligence-future-combat-dirk-hoke/.

[134] It is worth noting that Airbus Defence and Space has also sought to incorporate socio-technical design elements into its weapon systems to correspond with the NATO Human Views framework, which is discussed in the multilateral section. See: Robert A. Sharples, "Introduction of Human Views into Operational Capability Development within an Architectural Framework – March 2014," *MESAS 2014: Modelling and Simulation for Autonomous Systems* (Rome: May 5–6, 2014).

[135] Appendix VIII provides one version of analysis of how these trustworthy principles overlap and are in tension with the defense sector.

[136] Massimo Azzano and Sebastien Boria et al., "White Paper: The Responsible Use of Artificial Intelligence in FCAS – An Initial Assessment" (FCAS Forum, 2021), http://www.fcas-forum.eu/en/articles/responsible-use-of-artificial-intelligence-in-fcas.

[137] Azzano et al., "White Paper."

[138] It is worth noting that the White Paper "initial assessment" expanded on the question of targeting. They found that explainability and accountability would be important aspects to track going forward. Subsequent assessments on other case studies could be more directly relevant to AI-enabled decision support. Here, it is worth noting that Germany is in the lead for the unmanned components of FCAS, which may also be one reason that the FCAS Forum is focused on autonomy in weapons, further to the political reasons described in this section. See: Azzano et al., "White Paper"; Bundesverband der Deutschen Luft- und Raumfahrtindustrie e.V., "Das Future Combat Air System: Übersicht," 2021, 6, https://www.bdli.de/sites/default/files/2021-06/Übersicht%20FCAS.pdf.

[139] Azzano et al., "White Paper," 2021.

[140] Azzano et al., "White Paper," 2021.

[141] Michael Horowitz and Paul Scharre, "AI and International Stability: Risks and Confidence-Building Measures" (Center for a New American Security, January 12, 2021), https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[142] John R. Deni, *Security Threats, American Pressure, and the Role of Key Personnel: How NATO's Defence Planning Process is Alleviating the Burden-Sharing Dilemma* (Carlisle Barracks: U.S. Army War College Press, October 9, 2020), 1.

[143] The two-pronged approach to "dynamic adoption" and "responsible use" of AI has been championed by the author of the NATO White Papers on AI and autonomy (which are not in circulation). See: Edward Hunter Christie, "Artificial Intelligence at NATO: dynamic adoption, responsible use," *NATO Review,* November 24, 2020, https://www.nato.int/docu/review/articles/2020/11/24/artificial-intelligence-at-nato-dynamic-adoption-responsible-use/index.html.

[144] Christie, "Artificial Intelligence at NATO."

[145] North Atlantic Treaty Organization, "Emerging and disruptive technologies," last updated June 18, 2021, https://www.nato.int/cps/en/natolive/topics_184303.htm?selectedLocale=en.

[146] Mercier, "SACT's opening remarks to the NAC/MC Away Day."

[147] Mercier, "SACT's opening remarks to the NAC/MC Away Day."

[148] *NATO: READY FOR THE FUTURE: Adapting The Alliance (2018-2019)* (Brussels: North Atlantic Treaty Organization), November 29, 2019, 17, https://www.nato.int/nato_static_fl2014/assets/pdf/pdf_2019_11/20191129_191129-adaptation_2018_2019_en.pdf and D.F. Reding and J. Eaton, *Science & Technology Trends 2040: Exploring the S&T Edge* (Brussels: NATO Science & Technology Organization, 2020), 2.

[149] The following month, NATO also identified concrete priority areas—another important piece of the implementation map for the Alliance. See: North Atlantic Treaty Organization, "Emerging and disruptive technologies."

[150] Vivienne Machi, "Artificial intelligence leads NATO's new strategy for emerging and disruptive tech," *C4ISRNet*, March 14, 2021, https://www.c4isrnet.com/artificial-intelligence/2021/03/14/artificial-intelligence-leads-natos-new-strategy-for-emerging-and-disruptive-tech/.

[151] Machi, "Artificial intelligence leads NATO's new strategy for emerging and disruptive tech"; Sebastian Sprenger, "NATO tees up negotiations on artificial intelligence in weapons," *C4ISRNet*, April 27, 2021, https://www.c4isrnet.com/artificial-intelligence/2021/04/27/nato-tees-up-negotiations-on-artificial-intelligence-in-weapons/.

[152] Reding and Eaton, *Science & Technology Trends 2040*, 15.

[153] It is worth noting that the NATO Industrial Advisory Group has taken on a larger role in recent years.

[154] Michael Boardman and Fiona Butcher, *An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It* (Neuilly-sur-Seine: NATO Science & Technology Organization, October 10, 2019).

[155] Many allies still see network-enabled/network-centric operations as the predominant operational concept that guides their doctrinal, organizational, and acquisition-related adoption of technology.

[156] NATO Research and Technology Organization, *Human Systems Integration for Network Centric Warfare* (Neuilly-sur-Seine: February 2010), 2-3.

[157] NATO Research and Technology Organization, *Human Systems Integration for Network Centric Warfare*, 1-7.

[158] Boardman and Butcher, *An Exploration of Maintaining Human Control in AI Enabled Systems*, 3.

[159] Boardman and Butcher, *An Exploration of Maintaining Human Control in AI Enabled Systems*, 8–9.

[160] NATO Standardization Office, *NATO Standard ATrainP-2: Training in the Law of Armed Conflict Edition B Version 1* (Brussels: June 2019), 1.

[161] The NSCAI has also made similar connections, treating interoperability as a sixth pillar to the DOD principles framework because of its importance to safety when mapping out implementation and international engagement. See: National Security Commission on Artificial Intelligence, *Third Quarter Recommendations* (Washington, DC: 2020), 222–224.

[162] JAIC Public Affairs, "JAIC facilitates first-ever International AI Dialogue for Defense," Joint Artificial Intelligence Center, September 16, 2020, https://www.ai.mil/news_09_16_20-jaic_facilitates_first-ever_international_ai_dialogue_for_defense_.html.

[163] "Ministry of Defence, Singapore (MINDEF)," Facebook page, post on March 25, 2021, https://www.facebook.com/mindefsg/.

[164] National Security Commission on Artificial Intelligence, *Third Quarter* Recommendations, 222–224.

[165] For reference, 21 of the 27 EU member states are also NATO allies.

[166] Gabriela Baczynska, "U.S. among first foreign countries to join EU defence project, diplomats say," Reuters, May 5, 2021, https://www.reuters.com/world/us/us-among-first-foreign-countries-join-eu-defence-project-diplomats-say-2021-05-05/.

[167] Edward Hunter Christie, "The NATO Alliance and the Challenges of Artificial Intelligence Adoption" in *NATO Decision-Making in the Age of Big Data and Artificial Intelligence* (NATO Allied Command Transformation, the University of Bologna and Istituto Affari Internazionali, March 2021), 89.

[168] Finland's Presidency of the Council of the European Union, *Food for Thought Paper by Finland, Estonia, France, Germany, and the Netherlands: Digitalization and Artificial Intelligence in Defence* (Brussels: European Union, May 17, 2019), https://eu2019.fi/documents/11707387/12748699/Digitalization+and+AI+in+Defence.pdf/151e10fd-c004-c0ca-d86b-07c35b55b9cc/Digitalization+and+AI+in+Defence.pdf.

[169] High-Level Expert Group on Artificial Intelligence, "The High-Level Expert Group on Artificial Intelligence: Ethics Guidelines for Trustworthy AI," European Commission, April 8, 2019, 16.

[170] Relevant EU actors include: the European Commission for industrial policy (with separate portfolios on defense funding and AI); the European Council (with the Political and Security Council and additional working groups focusing on AI); the Strategic Policy Planning Division, EU Military Staff, and Security and Defence Policy groups inside the European External Action Service (respectively for policy, operational matters, and strategic coordination); and the European Defence Agency (for taxonomical work and cooperation). As discussed above, the European Parliament also exerts influence, though would not have ownership over this portfolio per se. At present, each of these actors handles a different aspect of the military AI portfolio, and each has its own relations with NATO counterparts, though with varying degrees of coordination.

[171] "Regulation (EU) 2018/1092 of the European Parliament and of the Council of 18 July 2018 establishing the European Defence Industrial Development Programme aiming at supporting the competitiveness and innovation capacity of the Union's defence industry," (Brussels: European Union, August 7, 2018), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32018R1092&from=EN.

[172] European Parliament Committee on Legal Affairs, *Guidelines for military and non-military use of Artificial Intelligence*, European Parliament, January 20, 2021, https://www.europarl.europa.eu/news/en/press-room/20210114IPR95627/guidelines-for-military-and-non-military-use-of-artificial-intelligence.

[173] New Zealand, which is also not in the PfD, has not issued a public approach but has different models to pluck from within the Five Eyes context.

[174] Zoe Stanley-Lockman, "Toward a Military AI Cooperation Toolbox: Modernizing S&T Defense Partnerships for the Digital Age" (Center for Security and Emerging Technology, forthcoming).

[175] Jin-Hee Cho, Patrick M. Hurley, and Shouhuai Xu, "Metrics and Measurements of Trustworthy Systems," Proceedings of *MILCOM 2016 - 2016 IEEE Military Communications Conference* (Baltimore, MD: November 1–3, 2016), 1237–1242.

[176] The NSCAI notes that trustworthiness is one of the focus areas of the TTCP Strategic Challenge on AI. Limited information is available about the Law & Ethics working group, except that the lead author of the Australian method document is a member. National Security Commission on Artificial Intelligence, *First Quarter Recommendations* (Washington, DC: 2020), 65; "The Ethics of AI

in Defence," Engineers Australia, June 15, 2021, https://www.engineersaustralia.org.au/event/2021/05/ethics-ai-defence-36781.

[177] This intended use often relates to transferring tactical control of AI-enabled systems between units, services, and nations. See: Robert S. Bolia, "The TTCP AI Strategic Challenge," *Fourth Annual Workshop on Naval Applications of Machine Learning* (San Diego, CA: February 2020), 24–7.

[178] Devitt et al., *A Method for Ethical AI in Defence*, 2020, 14–27, 33; "Pragmatic Tools for Considering and Managing Ethical Risks in AI for Defence – Detailed."

[179] For more information on the Allied Impact C2 system, see: Braulio Coronado, Crisrael Lucero, and Douglas S. Lange, "Discretizing and Managing the Task Environment," NATO Science & Technology Organization, April 11, 2019, 3–8, https://www.sto.nato.int/publications/STO Meeting Proceedings/Forms/AllMPs.aspx?RootFolder=%2Fpublications%2FSTO Meeting Proceedings%2FSTO-MP-HFM-300&FolderCTID=0x0120D5200078F9E87043356C409A0D30823AFA16F602008CF184CAB7588E468F5E9FA364E05BA5&View=%7B72ED425F-C31F-451C-A545-41122BBA61A7%7D; Mark Draper, Allen Rowe, and Jessica Bartik, "TTCP Autonomy Strategic Challenge," U.S. Air Force Research Laboratory, 2019, https://nari.arc.nasa.gov/sites/default/files/attachments/Day 1 MarkDraper Slides.pdf. See also: Stanley-Lockman, "Toward a Military AI Cooperation Toolbox," forthcoming.

[180] Cho et al., "Metrics and Measurements of Trustworthy Systems," 1237.

[181] The metrics are: "(1) a system of concern and its features, states and behavior; (2) threats, including faults, errors, and failures caused by deliberate actions (i.e., attacks) or non-deliberate actions; (3) key metrics of trustworthiness; (4) means to build trustworthy systems; (5) relationships between assessment (e.g., red teaming, vulnerability assessment, penetration testing) and submetrics (or attributes) of a trustworthiness metric." Cho et al., "Metrics and Measurements of Trustworthy Systems," 1238.

[182] More specifically, these characteristics are "hardware; software; the network; the effect of human factors (i.e., users or system designers/analysts); and physical environments." Cho et al., "Metrics and Measurements of Trustworthy Systems," 1238.

[183] In addition to deploying alongside each other, this could conceivably include sharing data inputs like models or training data for co-developed or emulated systems. Understanding the limitations of an ally's capability would be important for development and deployment alike.

[184] Devitt et al., *A Method for Ethical AI in Defence*, 49–50.

[185] Margarita Konaev, Tina Huang, and Husanjot Chuhal, "Trusted Partners: Human-Machine Teaming and the Future of Military AI" (Center for Security and Emerging Technology, February 2021), 18–23, https://cset.georgetown.edu/publication/trusted-partners/.

[186] Other adjacent forms of cooperation could include, for example, how allies increase access to one another's defense sectors by implementing export control exemptions or incentivizing transfers of non-controlled parts.

[187] Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense*.

[188] Christie, "Artificial Intelligence at NATO," 2020.

[189] Kathleen J. McInnis, "Lessons in coalition warfare: Past, present and implications for the future," *International Politics Review* 1 (2013): 78–90; Edward Hunter Christie, "Artificial Intelligence at NATO: dynamic adoption, responsible use."

[190] Defence Ethics Committee, *Opinion on the Integration of Autonomy into Lethal Weapon Systems*, 5–6.

[191] The Australian Method document provides examples of AI uses that relate to maintaining freedom of action (under "force protection"), including: "AI in Cyber Network Defense," "AI used to develop and employ camouflage and defensive deception systems and techniques," "AI to identify potential vulnerabilities in a friendly force that requires protection," and "AI used to simulate potential threats for modelling and simulation or rehearsal activities," among others. French language is more explicit about interoperability as part of its guidelines for controlled AI, but does not provide as much detail. See: Devitt et al., "A Method for Ethical AI in Defence," 58; AI Task Force, *Artificial Intelligence in Support of Defence*, 9.

[192] U.S. Department of Defense, "DOD Adopts Ethical Principles for Artificial Intelligence," February 24, 2020; Office of the Deputy Secretary of Defense, *Implementing Responsible Artificial Intelligence in the Department of Defense* (Washington, DC: May 26, 2021), 2; Organization for Economic Co-operation and Development, Recommendation of the Council on Artificial Intelligence, adopted on May 21, 2019; Defense Innovation Board, *Supporting Document*, 27–31.