

Beyond Alliance: Assessing the Utility of NATO's Principles of Responsible Use for AI in Defence for States Outside the Alliance

Daniel Trusilo

University of St. Gallen
daniel.trusilo@unisg.ch

Abstract

NATO's Principles of Responsible Use for AI in Defence (PRUs), which cover lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation, sit at the foundation of a growing network of alliance structures that shape how member states develop, fund, and field dual-use AI systems. How the PRUs travel beyond NATO members, and what their adoption yields for other states, remains under-examined. Drawing on primary sources and expert interviews, this paper maps four structures that operationalise the PRUs: Allied Command Transformation's Innovation Branch, the NATO Innovation Fund, the Data and AI Review Board, and the Defence Innovation Accelerator for the North Atlantic. It then sets NATO's approach against two well-developed national alternatives, the United States Department of War's AI Ethical Principles and the United Kingdom's Ambitious, Safe, Responsible framework. Using Switzerland as an illustrative case of a non-member state with a competitive AI and robotics ecosystem and a longstanding policy of armed neutrality, the paper draws out what an adopting state stands to gain or lose by engaging with the PRUs. The work speaks to ongoing debates on cross-national AI governance, dual-use technology adoption, and the operationalisation of ethical principles in real institutional settings.

Introduction

Artificial Intelligence (AI), in the form of both software and robotic systems, is rapidly altering the global security environment (Sayler 2020). This rapid change is exacerbated by the dual-use nature of AI in which systems designed by civilian entities for civilian purposes are applied to security contexts. The rapidly evolving ecosystem of emerging technology being applied to security contexts is also a fundamental aspect of a geopolitical climate of increasing great-power competition (Soare 2021). These factors are impacting the way socio-technical systems are developed, adopted,

and operationalised, with questions about the ethical values of stakeholders often a secondary consideration in the face of what may be portrayed as both a threat to security and strategic opportunity.

In the absence of international standards related to dual-use AI, NATO's Principles of Responsible Use for AI in Defence (PRUs), represent a recognised set of normative values that can be applied to facilitate the ethical development, adoption, and operationalisation of such systems. As an alliance of thirty-two member countries, NATO's approach to emerging and disruptive technology (EDT) has the potential to broadly influence military applications of such technology (Herzog and Kunertova 2024). Expanding on this idea, Stanley-Lockman and Trabucco noted in 2024, that NATO is positioned to not only define norms but also develop standards that will impact the effectiveness of militaries. Though it remains to be seen exactly how the PRUs will be operationalised by NATO member states, their consideration is being woven into a network of structures, both inside NATO and adjacent to it, that govern how alliance members incorporate and leverage EDTs.

Adopted in 2021 and reaffirmed in the alliance's revised AI strategy in 2024, the PRUs cover lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation (NATO 2021; NATO 2024). For member states, the network of structures that are using the PRUs sets a basis in terms of access to funding, test infrastructure, and interoperability standards. For non-member states with stakes in dual-use AI, the network functions differently, presenting possibilities for engagement contingent on partner arrangements or alignment with shared standards. This raises a question the existing literature on responsible military AI and dual-use AI has largely set aside: what does the PRU framework offer, and what does it cost, to a state that is not a NATO member?

This paper contributes to the growing body of literature on the ethics of dual-use AI by first mapping the four NATO structures most central to operationalisation of the PRUs: Allied Command Transformation (ACT) Innovation Branch, the NATO Innovation Fund (NIF), the Data and AI Review Board (DARB), and the Defence Innovation Accelerator for the North Atlantic (DIANA). Second, it sets NATO's approach against two well-developed national alternatives: the United States Department of War's AI Ethical Principles and the United Kingdom's Ambitious, Safe, Responsible framework. Third, it uses Switzerland, a non-NATO state with a competitive AI and robotics ecosystem and a longstanding policy of armed neutrality, as an illustrative case of how a state can weigh engagement with the PRUs as it develops its own approach to dual-use AI. By taking this approach, this paper maps a network of institutions that has not been systematically described in the AI ethics literature and shows, through one case, what an adopting state stands to gain or lose. The Swiss case is a most-favorable case for this analysis as it is a financially and technologically well-resourced, politically stable, non-member state. Additionally, Switzerland holds a defense posture grounded in armed neutrality and is home to firms and research groups that already engage with NATO members through civilian and academic channels (Trabucco and Maas 2023). If it is too complicated or difficult for Switzerland to engage with the PRUs it is likely more so for states with fewer resources and/or greater political instability.

Methodological Approach

To gain insights on the structures related to NATO's PRUs and how they are being operationalised in practice, this work draws on primary documents and eight semi-structured expert interviews across four institutional settings conducted between September and October 2025. Interviewees were selected through purposive sampling to cover the range of institutions examined (e.g., current NATO officials involved in the DARB, a senior contributor to the UK's approach, officials involved with the American approach, and independent researchers working on responsible military AI). Interviewees were granted anonymity given the sensitivity of their roles, and are identified in the text by role and month only. Interviews were documented through contemporaneous notes and analysed against the document record. Where possible, interview claims were cross-checked against public NATO documents or corroborated across more than one interviewee. Documentary sources include the unclassified summaries of NATO's 2021 and 2024 AI strategies, the 2022 Strategic Concept, NATO Parliamentary Assembly reports, and public materials from the US and UK defence ministries.

The paper proceeds as follows. The second section sets out the PRUs and the four structures that carry them. The

next section maps the comparative landscape, drawing on the US and UK approaches. The paper then takes up the Swiss case. A discussion then draws out what the case suggests about the wider question of how the PRUs travel. The final section concludes.

The PRUs and Their Institutional Embedding

In October 2021 NATO adopted the AI Strategy (henceforth *The Strategy*) and launched the Defence Innovation Accelerator for the North Atlantic (DIANA) (NATO 2021). According to the unclassified summary of *The Strategy* and multiple expert interviews, the PRUs are foundational to NATO's development and operationalisation of EDTs. Based on widely accepted ethical, legal, and policy commitments, NATO's PRUs are intended to go beyond a normative ethical code, facilitating interoperability among allies.

Related to *The Strategy*, NATO's current Strategic Concept emphasises the changing security environment placing emphasis on promoting innovation and increasing investments in EDTs as a matter of maintaining military advantage (NATO 2022). Along with the strategic concept, the alliance also established a Data and AI Review Board (DARB) to "bring together ethicists, policy experts, engineers, developers and operators" to operationalise the PRUs and develop a "NATO Responsible AI certification standard" (NATO 2023). The underlying motivation for these efforts is captured in a 2024 special report on AI prepared for the NATO Parliamentary Assembly, which states: "The military use of AI is neither inherently 'good' or 'bad'. What matters is the context and the rules that guide the application of AI" (Clement 2024). To address the rapidly expanding capabilities of AI and update the outcomes identified in 2021, *The Strategy* was revised in July 2024 (NATO 2024). The PRUs were not altered under the revision and still apply to NATO AI use-cases across the full-lifecycle.

NATO's AI initiatives indicate recognition that emerging AI technologies are profoundly altering many aspects of defense and warfighting. This aligns with Onderco's (2025) analysis of the Ukraine conflict that reached three conclusions of direct relevance: the boundary between civilian and military sectors has narrowed sharply; the war has pulled many aspects of civilian life into military use; and exploration of autonomy in military systems has accelerated to the point that adoption is now a matter of choice more than capability. This recognition is paired with an emphasis on incorporating normative principles related to the responsible-use of AI for defense, which can be seen as intended to shape the use of AI for defense toward liberal democratic values and respect for human rights (Herzog and Kunertova 2024).

The PRUs do not stand alone. The alliance has built and adapted a set of institutions that turn the principles from a declaration into a set of working processes. The next four

subsections will briefly highlight the specific structures that have been created in order to broadly support the adoption of EDTs in accordance with the PRUs.

Each of the four structures examined below sits at a different point in the technology adoption pathway, from early-stage funding to certification to operational testing. Part of the motivation behind these structures is that the legacy procedures NATO uses for buying, testing, and fielding equipment cannot keep pace with the rate at which AI capabilities now develop (Clement 2024). The newer structures aim to close that gap. Some of these institutions are adjacent to NATO rather than fully internal. The NATO Innovation Fund, for example, is an independent venture fund supported by twenty-four allies; DIANA is a NATO body but draws on a network of accelerator sites and test centres owned by allies. The PRUs run through all four as a shared reference point, even where the principles do not appear by name in public materials.

Allied Command Transformation's Innovation Branch

NATO's Allied Command Transformation (ACT) has multiple initiatives related to the adoption of EDTs. In 2012, ACT's Innovation Branch was established with the objective of leading NATO's digital transformation. The Innovation Branch has evolved to focus on open military innovation with the objective of bringing together end-users, providers, and capability designers to leverage EDTs through a NATO Innovation Network (ACT 2025a). In 2024, building on the Innovation Branch's validated projects, ACT initiated the Innovation Continuum (ACT 2024). The Innovation Continuum is a year-long program involving academia, industry, and allies with the objective of providing a platform for the experimentation and demonstration of EDTs from deep-tech startups to government research agencies (ACT 2025b). The 2025 iteration concluded with a week-long event in Istanbul where participants demonstrated technologies of direct relevance to the AI and robotics field, including counter-drone technologies as well as heterogeneous uncrewed systems operating collaboratively across multiple domains (ACT 2025c). There is no direct reference to the PRUs in the ACT's press releases, but the PRUs are considered a baseline for NATO's AI strategy, which is also fundamental to driving ACT's efforts.

The NATO Innovation Fund (NIF)

The NIF is a one-billion-euro multi-sovereign venture fund with a fifteen-year horizon, established in June 2022 with contributions from twenty-four allies (NIF 2025). It exists to draw private capital into defence and dual-use technology in allied countries and to help the supported technologies reach allied defence ministries.

Fundamental to the Fund's investment objectives are the NATO PRUs as ethical principles are seen as critical to balancing innovation with trust, accountability, and responsible use standards (Murray, Schneider-Sikorsky, and Latham 2025). That said, though the NIF is interested in the work of the DARB and its sub-groups, there is currently no specific output from the DARB that the NIF can apply (Interview with NATO representative, October 2025). The NIF's emphasis on ethical and responsible technology is not a case of ethics washing but rather a strategic move to address exclusions to defense spending by European financial institutions in accordance with stated environmental, social, and governance considerations (ESG). The Innovation Fund does this through a Responsible Use Framework that analyses: if technologies are designed and deployed with the PRUs and can be trustworthy to both operators and the public, whether a company has adequate accountability and governance, if risks are mitigated through appropriate safeguards, and by monitoring and guiding companies as they scale and innovate to ensure continuity (Murray, Schneider-Sikorsky, and Latham 2025).

In contrast to the prevailing perceptions toward the defense industry, the NIF views financial support of private sector companies developing principled defense technology as fully aligned with ESG objectives stating: "ESG and defense should be seen as complementary, requiring a more nuanced understanding of the sector's critical role in fostering long-term societal peace and stability" (Murray, Schneider-Sikorsky, and Latham 2025). But such a position is only valid if the supported private sector companies are aligned with the implied moral and ethical values. Therefore, the PRUs offer a normative foundation for the NIF to build on.

The Data and AI Review Board (DARB)

The DARB is intended to be an interdisciplinary forum composed of computer scientists, engineers, lawyers, and philosophers nominated by Allies, which serves as the focal point for NATO's efforts to operationalise NATO's six PRUs. Though there is a general convergence on the PRUs, developing certification standards according to these normative principles has been challenging. Therefore, the DARB is focused on developing processes using the PRUs as a heuristic (Interview with NATO official, October 2025).

In a typical DARB meeting, twenty to twenty-three representatives attend across two sub-groups (Interview with NATO official, October 2025). One sub-group is focused on standards, assessment, and tools while a second sub-group is focused on knowledge sharing. Allies may choose to use the DARB's outputs, such as the RAI toolkit, on a voluntary basis (NATO 2022b). Currently, NATO's RAI toolkit, aimed at the internal NATO enterprise, is a replication of the publicly available US Department of War's (DoW/DoD) Chief Digital and Artificial Intelligence Office (CDAO)

RAI toolkit (Interview with senior researcher, October 2025). Because NATO's version of the RAI toolkit is difficult to access outside of the NATO enterprise, it is not incorporated into normal workflows and has had limited adoption. However, there is an effort underway to refine and revise the approach. The DARB's standards, assessment, and tools subgroup is aspiring to find a more flexible approach that is contextually relevant and respectful of existing NATO processes while also taking into account multiple allies' approaches to RAI (Interview with NATO official, October 2025).

The Defence Innovation Accelerator for the North Atlantic (DIANA)

DIANA is the alliance's startup accelerator for dual-use deep tech (NATO 2025). It runs public challenges that civilian startups can apply to address, and offers selected innovators grants together with access to sixteen accelerator sites and over two hundred test centres across allied territory. DIANA does not directly engage with the DARB's operationalisation of the PRUs, though it keeps informed (Interview with NATO official, October 2025). The PRUs nonetheless reach DIANA participants indirectly, since the NATO Innovation Fund, which often co-invests, uses the PRUs as part of its Responsible Use Framework.

The PRUs as a Connective Layer

Taken together, the four structures show that the PRUs work as a connective layer across NATO's wider effort to bring in EDTs including dual-use AI. The principles are not embedded in any single piece of binding policy. They are instead integrated through a set of processes that allies and partners encounter at different points: when applying for funding, when seeking access to test centres, when contributing to standards, and when shaping doctrine. The next section sets this institutional reach against two well-developed national alternatives.

Comparative Landscape: US and UK Approaches

NATO is not the only source of principles for responsible military AI. A number of states have set out their own (Stanley-Lockman 2021). Two are well-developed enough to serve as alternative models: the United States and the United Kingdom.

US Department of War's AI Ethical Principles

The US Department of War (formerly Department of Defense, DoD) adopted five AI Ethical Principles in February 2020: responsible, equitable, traceable, reliable, and governable (DoD 2020). The principles apply to both combat and

non-combat use-cases, and serve as the foundation for the publicly available CDAO Responsible AI toolkit. That toolkit gathers about seventy open-source tools applicable across the AI lifecycle, presented in a modular form that lets users skip steps that do not apply to the system at hand (CDAO 2024). Because the CDAO toolkit is openly accessible and modular, it has likely seen more uptake than NATO's mirror version, which is harder to reach and currently lacks the skip-step function (Interview with senior researcher, October 2025).

DoD Directive 3000.09 on Autonomy in Weapon Systems, reissued in 2023, sits alongside the principles and offers the first directive of its kind globally (DoD 2023). It was reflected in US foreign policy through the Political Declaration on Responsible Military Use of AI and Autonomy, introduced at the 2023 REAIM Summit, to which over fifty states, Switzerland among them, are signatories (DoS 2023; Horowitz 2024).

United Kingdom's Ambitious, Safe, Responsible Framework

The UK's Ethical Principles for AI in Defence cover human-centricity, responsibility, understanding, bias and harm mitigation, and reliability (MoD 2022). The principles were published in the 2022 policy paper *Ambitious, Safe, Responsible*, and operationalised in 2024 through Joint Service Publication 936 as part of the UK Defence AI Strategy (MoD 2024).

The UK approach rests on the view that responsible-AI processes raise assurance between allies while helping teams anticipate barriers to adoption. The published instruments take the form of minimum viable products that can be tested and refined: a risk management framework, AI assurance question sets, and case studies. A senior contributor to the UK approach noted in interview that the national commitment to operationalising the principles is strong, and that a national approach can in some respects be more thorough than a multi-lateral one (Interview with senior contributor to UK approach, October 2025). The UK principles apply across the full range of AI applications in defense, combat and non-combat alike.

What the Comparison Shows

The three approaches converge on much of the same ground. All three name principles around responsibility, reliability, governability or its near-equivalents, and either bias mitigation or harm mitigation. Where they differ is in institutional reach. The US approach binds one ministry, with an openly accessible toolkit. The UK approach binds one ministry with a published implementation framework and an explicit set of working instruments. NATO's approach is alliance-wide, woven through a set of structures that span funding, certification, and capability development, but rests on voluntary

uptake by members and reaches non-members only by indirect routes such as partner arrangements.

The next section uses Switzerland as a case to analyse what adopting any of these sets of principles means in terms of potential opportunities and risks.

Switzerland: An Illustrative Case

Switzerland is a most-favourable case for this analysis for multiple reasons. First, it sits outside NATO and holds a defence posture grounded in armed neutrality, a position shaped over more than two centuries and written into its current foreign-policy doctrine. Second, it has a globally competitive AI and robotics ecosystem, anchored in research institutions, deep-tech firms, and a federal procurement agency that works at the nexus of civilian and defence applications. Third, Switzerland offers a politically stable context that is well-resourced both financially and technologically, making it possible to make substantive policy decisions that can be followed up with action. Switzerland thus represents a most-favourable case for a class of states that face the question of how to engage with the growing body of responsible AI frameworks, including NATO's PRUs, in an increasingly unstable geopolitical environment and rapidly developing dual-use AI technology. If Switzerland finds it too complicated to engage with NATO's PRUs it is likely that a state with less political stability, exposure to ongoing conflict, or fewer financial and technological resources will find engagement with the PRUs more difficult.

This section sets out the considerations a state in Switzerland's position would face. It organises those considerations into four open questions: what a state stands to gain from access to NATO's wider structures, what it gains and gives up on standards and influence, how alignment fits with normative commitments and strategic autonomy, and what it stands to lose by staying outside.

Access and Resources

The first question is access. The NATO Innovation Fund offers capital and market access to deep-tech startups from contributing allies. For a Swiss firm, direct participation in the Fund would require a Swiss contribution or some other route through which the country becomes eligible. The Fund applies the PRUs through its Responsible Use Framework, so engagement with the Fund implies alignment with the principles, whether or not Switzerland adopts them formally. A concrete first step for a state in Switzerland's position is to establish the specific requirements that domestic government and private-sector actors would need to meet to take part.

Access also runs through DIANA and ACT. DIANA's challenges are open to startups from allied countries, and

DIANA-affiliated test centres span the alliance. The Innovation Continuum offers a forum in which Swiss firms could demonstrate capabilities alongside allies. The PRUs reach these venues less directly than they reach the Fund, but they shape what counts as an acceptable system within the wider ACT effort.

Resource access is not the only consideration on this axis. Smaller states tend to gain disproportionately from alliance-wide arrangements because such arrangements help close capability gaps and build ties across the public, private, and academic sectors (Herzog and Kunertova 2024). The transatlantic defence primes do not currently have the AI and robotics talent needed to build the kinds of systems now reshaping the security environment (Herzog and Kunertova 2024). That gap creates room for innovative solutions from smaller ecosystems if access channels are open.

Standards, Interoperability, and Influence

The second question relates to both ethical and technical standards. NATO's Standardization Agreements (STANAGs) bind allies to common technical specifications across fuels, communications, and increasingly AI and robotic systems. STANAG 4586, for example, sets standard interfaces for UAV control systems, and the alliance has begun drafting new STANAGs that will cover AI testing and evaluation (Marques 2012; NATO 2023). Such standards have far-reaching effects: JP-8 jet fuel, set originally as a NATO standard, has since seen global uptake.

A non-member state can engage with this standards landscape in two ways. It can adopt the standards as a downstream user, which lowers the transaction costs of future co-operation but does not give it a voice in setting them. Or it can engage upstream, contributing to the working groups that shape the standards. The latter route is open to non-members through partner arrangements. According to a NATO representative interviewed for this work, the DARB sub-groups are open to non-member state participation, e.g. Swiss participation, if certain agreements are in place (Interview with NATO official, October 2025). A concrete first step for Switzerland would be to designate a representative to take part in the DARB sub-groups, including the standards, assessment, and tools sub-group and the knowledge-sharing sub-group. The representative would need a strong grasp of responsible-AI best practice and security operations, and could come from government, the military, or academia.

A coalition of states larger than the few advanced military powers may also be able to shape the outcome of conversations on the use of AI in military applications (Canfil and Kania 2024). New patterns of cooperation around the military use of AI support this possibility (Trabucco and Maas 2023). For Switzerland, engagement at the standards level

offers a way to contribute to this coalition without binding the country to collective defence.

Normative Alignment and Strategic Autonomy

The third question is normative. NATO's approach to EDTs places ethical and normative commitments at the foundation of its strategy (NATO Advisory Group 2021). As the NATO Secretary-General put it in 2021, NATO can be the platform to address ethical aspects of these technologies, including by developing guiding standards (NATO 2021b). For a state with a long tradition of grounding its foreign policy in international law and humanitarian principles, alignment with the PRUs has surface appeal. The principles draw on shared commitments and push in a direction that fits Swiss positions on questions like distinction and proportionality in armed conflict.

Yet alignment carries costs. NATO's PRUs apply to systems used for military purposes, while the EU AI Act, which applies in part to Swiss firms through market access, specifically excludes military applications. In practice, this means a Swiss firm building a dual-use system faces overlapping and at times conflicting regulatory expectations potentially resulting in the same system being governed twice, or not at all, depending on the use-case. For example, a civilian AI human resource management system that is deemed "high-risk" under the EU AI Act, would be subject to the Act's regulations and, if adopted for military applications by NATO HQs, also subject to any practices the DARB deemed necessary to operationalise the PRUs. Or, in the event that a private-sector actor develops an AI system for strictly military use-cases, such as an image recognition system for military intelligence analysis, the EU AI Act is not applicable. However, such a system could be deployed by the military of a NATO member state without any consideration of the PRUs as how and when the PRUs are operationalised by member states (or partner nations) is voluntary.

This conflicting regulatory environment is further complicated by the absence of a shared understanding of the PRUs across allies, which raises questions about both ethical and technical interoperability: how does an adopting state implement the principles in line with its own national requirements while keeping a standard form of implementation (Interview with NATO official, October 2025)?

For Switzerland specifically, alignment must also be weighed against the policy of armed neutrality. Long-standing federal practice holds that Swiss neutrality permits co-operation with NATO on standards and on the exchange of information, but rules out alignment that commits the country to collective defence. Adoption of the PRUs would not by itself cross that line. But the further Switzerland engages with the institutions that carry the PRUs, the more attention the boundary needs.

Exclusion and Resource Vulnerabilities

The fourth question relates to any costs a state incurs for non-engagement. If Switzerland chooses not to align with the PRUs and not to engage with the structures that carry them, Swiss firms risk exclusion from the responsible-AI market the alliance is shaping. Other first movers, including states with little regard for the rules-based order, may drive the adoption of less safe or less ethical systems (Drexel and Kelley 2023). Allies adopting interoperable solutions may rule out Swiss systems on technical or normative grounds, even where the Swiss system would otherwise be a strong fit (Brunner 2022; Schmid, Brunner, and Taylor 2017). Export controls on chips and on the underlying compute infrastructure for AI add a further pressure. Access to these resources is essential for the Swiss robotics ecosystem to keep its position at the frontier (Shivakumar, Wessner, and Howell 2025). NATO has no role in setting export controls, but engagement with alliance institutions shapes which side of the controls a country sits on in practice.

A separate consideration runs in the other direction. Closer engagement with NATO opens channels through which Swiss talent and technical solutions can be recruited or adopted by allies. That has positive effects on the Swiss ecosystem in some respects, including capital, market access, and prestige, but also risks drawing capability away from the domestic enterprise. The balance is not easy to read in advance.

An Interdisciplinary Committee

Across the four questions above, the case for a structured national engagement with the PRUs is strong, but it depends on the country making concrete choices about how to apply each principle in its own context. Such a finding suggests that adopting states would benefit from the creation of an interdisciplinary committee that could map the PRUs to national laws and governance structures with the intent to concretely identify the benefits and risks of each of the principles. A state that wanted to take on this task could, for example, establish a committee drawn from at least two technical experts in robotics or computer science; two military personnel with future deployment responsibilities; two legal experts familiar with the Law of Armed Conflict, Article 36 weapons reviews, and domestic law; and two private-sector representatives. The committee would consider each principle in context. The Principle of Governability is a useful example: it is defined as the capacity for appropriate human-machine interaction, the detection and avoidance of unintended consequences, and the ability to disengage or deactivate a system when needed (NATO 2021). What counts as appropriate human-machine interaction is itself contested across the alliance (Stanley-Lockman 2021). Therefore, if

we consider the case of Switzerland, adoption of the principle would require a domestic definition that fits Swiss legal and operational practice.

Discussion

The Swiss case suggests three wider points about how the PRUs travel beyond NATO members.

First, the principles travel more easily than the institutions that carry them. A non-member state can in principle align with the PRUs as a set of normative commitments without ever engaging with NATO. The principles draw on shared ethical, legal, and policy traditions and overlap substantially with the US and UK approaches. This is not at all surprising as the US and UK are both founding members of NATO and have played a role in the development of the PRUs. Therefore, a state that adopts a domestic set of responsible-AI principles modelled on the PRUs gains the normative benefits, including international signalling and a stable reference point for domestic regulation, at low institutional cost.

Second, the structures travel less easily. Access to the NATO Innovation Fund, DIANA, and the DARB sub-groups depends on either alliance membership or a partner arrangement. A non-member state can read about these structures, but cannot draw on them in the way an ally can. The institutional reach of the PRUs is therefore a function of the state's relationship with the alliance, not of the principles themselves. This matters for any non-member state weighing alignment, since most of the practical resources the alliance offers flow through the structures rather than through the principles.

Third, the case for engagement rests less on the PRUs as principles and more on the wider responsible-AI ecosystem the alliance is directly and indirectly shaping. The Innovation Fund's Responsible Use Framework sets a *de facto* standard that European investors increasingly look to. The DARB's emerging toolkit, if made more accessible, may set a standard that allies and partners draw on. Non-engagement in these processes means a non-member state lets other actors set the terms on which its firms must compete. The question for a state is therefore not so much whether the PRUs are the right principles, but whether the cost of standing outside the structures that carry them is one the state is willing to bear.

Limitations

This paper takes the PRUs as a reference point rather than engaging with the growing body of literature related to the viability of ethical principles for AI in defence such as the work by Taddeo, McNeish, Blanchard, and Edgar (2023) or the friction caused by the subjective nature of normative values such as the challenge of ethical interoperability for AI in defence (Danks and Trusilo 2022). The institutional features

analysed here delineate where such evaluations would begin. Uptake of any formal set of normative values is voluntary with no mechanism to compel a member state to apply the principles to a given system. And it is clear that many key terms are contested, as the disputed meaning of appropriate human-machine interaction under the governability principle shows. The certification infrastructure that would be required to give the principles force does not currently exist. Therefore, it is necessary to ask whether the PRUs constrain the development and use of military AI, lend it normative cover, or serve to further the ethical development and operationalisation of technologies in a substantive way.

The Need for Additional Comparative Work

The Swiss case examined in this work is one illustration of one type of non-member state. Developed states that are strategic partners of NATO, such as Japan or Australia, would face a different set of considerations. States that are EU members but not NATO members, such as Ireland or Austria, would face still another set. Additionally, states engaged in active armed conflict, where operational reality often diverges from normative frameworks, will present different risks and opportunities related to dual-use AI and how they might assess the value of the PRUs. Lastly, developing-country contexts, where adoption of normative frameworks may deepen rather than close technological asymmetry, presents other risks and opportunities. Therefore, a comparative study across two or more of such cases would help test whether the patterns identified in this work hold more widely, and would sharpen understanding of which features of the PRUs are general to responsible-AI governance and which are specific to particular contexts and factors.

Conclusion

This paper has mapped four NATO structures that carry the Principles of Responsible Use for AI in Defence: ACT's Innovation Branch, the NATO Innovation Fund, the Data and AI Review Board, and the Defence Innovation Accelerator for the North Atlantic. It has set NATO's approach against the US and UK alternatives and used Switzerland as an illustrative case to draw out what an adopting state stands to gain or lose. The work shows that the PRUs function as a connective layer across a wider network of NATO institutions, and that the question of alignment for a non-member state turns less on the principles themselves than on the institutional structures that carry them.

For a state in Switzerland's position, the practical steps include sending a representative to the DARB sub-groups, setting up an interdisciplinary committee to consider each principle in the domestic context, establishing the requirements for participation in the NATO Innovation Fund, and

tracking the standardization work that will shape future systems. These steps do not commit the state to alignment with the PRUs in full. They do let the state engage with the structures that carry the principles, and through that engagement, shape both the standards related to the ethical development and operationalisation of dual-use AI systems and the conditions under which its firms will compete.

The work suggests a number of lines for further research. A comparative study across two or more non-member states would test whether the patterns identified here hold more widely. A longitudinal study of the DARB's evolving toolkit and of the Innovation Fund's portfolio would help track how the PRUs operationalise in practice. And a closer look at the interaction between NATO's military framing and the EU AI Act's civilian framing would help map the regulatory terrain that dual-use AI firms must navigate.

Acknowledgments

Many thanks to armasuisse Science + Technology and the University of St. Gallen Law School for supporting the research that served as the foundation for this article. I am also grateful to the NATO and Defence professionals and experts that shared their time and insights. Lastly, many thanks to Prof. Dr. Thomas Burri who provided valuable feedback on early drafts of this work.

References

- Allied Command Transformation (ACT). 2024. Innovating for Tomorrow: NATO Allied Command Transformation's Innovation Continuum 2024 Series Takes Off in La Spezia. 29 February. <https://www.act.nato.int/article/innovation-continuum-2024-takes-off-in-la-spezia/>. Accessed: 2025-09-20.
- Allied Command Transformation (ACT). 2025a. The Innovation Hub for Allied Command Transformation. <https://innovationhub-act.org/what-we-do/>. Accessed: 2025-09-20.
- Allied Command Transformation (ACT). 2025b. Innovation Continuum. <https://www.act.nato.int/activities/innovation-continuum/>. Accessed: 2025-09-20.
- Allied Command Transformation (ACT). 2025c. NATO's Innovation Continuum 2025 Culminates in SHINE Event. 17 October. <https://www.act.nato.int/article/shine-2025/>. Accessed: 2025-11-05.
- Brummer, M. 2022. Innovation and Threats. *Defence and Peace Economics* 33(5): 563–584. doi.org/10.1080/10242694.2020.1853984.
- Canfil, J. K., and Kania, E. B. 2024. Mapping State Participation in Military AI Governance Discussions. In *The Oxford Handbook of AI Governance*, edited by J. B. Bullock, 895–913. Oxford: Oxford University Press.
- CDAO (Chief Digital and Artificial Intelligence Office). 2024. RAI Tools List. <https://rai.acqbot.com/tools-list>. Accessed: 2025-10-08.
- Clement, S. 2024. NATO and Artificial Intelligence: Navigating the Challenges and Opportunities. NATO Science and Technology Committee. 24 November. <https://www.nato-pa.int/document/2024-nato-and-ai-report-clement-058-stc>. Accessed: 2025-09-05.
- Danks, D., and Trusilo, D. 2022. The Challenge of Ethical Interoperability. *Digital Society* 1(11). doi.org/10.1007/s44206-022-00014-2.
- DoD (US Department of Defense). 2020. DOD Adopts Ethical Principles for Artificial Intelligence. Washington, DC: Department of Defense. <https://www.war.gov/News/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/>. Accessed: 2025-09-01.
- DoD (US Department of Defense). 2023. Directive 3000.09: Autonomy in Weapon Systems. Washington, DC: Office of the Under Secretary of Defense for Policy. <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>. Accessed: 2025-09-01.
- DoS (US Department of State). 2023. Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy.
- Drexel, B., and Kelley, H. 2023. China Is Flirting with AI Catastrophe: Why Accidents Pose the Biggest Risk. *Foreign Affairs* May 2023.
- Herzog, S., and Kunertova, D. 2024. NATO and Emerging Technologies—The Alliance's Shifting Approach to Military Innovation. *Naval War College Review* 77(2): 47–70. <https://www.jstor.org/stable/48821934>.
- Horowitz, M. C. 2024. The State of DoD AI and Autonomy Policy. Interview by Gregory C. Allen, 9 January. Center for Strategic and International Studies. <https://www.csis.org/analysis/state-dod-ai-and-autonomy-policy>. Accessed: 2025-09-23.
- Marques, M. M. 2012. STANAG 4586—Standard Interfaces of UAV Control System (UCS) for NATO UAV Interoperability. NATO Standardization Agency. Almada, Portugal: Escola Naval. <https://publications.sto.nato.int/publications/STO%20Educational%20Notes/STO-EN-SCI-271/EN-SCI-271-03.pdf>. Accessed: 2025-11-02.
- MoD (UK Ministry of Defence). 2022. Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in Defence. London: MoD. https://assets.publishing.service.gov.uk/media/62a9b1d1e90e07039e31b8cb/20220614-Ambitious_Safe_and_Responsible.pdf. Accessed: 2025-11-10.
- MoD (UK Ministry of Defence). 2024. Dependable Artificial Intelligence (AI) in Defence Part 1: Directive, JSP 936 V1.1. London: MoD. <https://www.gov.uk/government/publications/jsp-936-dependable-artificial-intelligence-ai-in-defence-part-1-directive>. Accessed: 2025-11-10.
- Murray, F.; Schneider-Sikorsky, P.; and Latham, B. 2025. Responsible Investing in Defence, Security and Resilience. NATO Innovation Fund. https://www.nif.fund/wp-content/uploads/2025/01/NIF_Davos_v3.pdf. Accessed: 2025-11-09.
- NATO. 2021. Summary of the NATO Artificial Intelligence Strategy. 22 October. https://www.nato.int/cps/en/natohq/official_texts_187617.htm. Accessed: 2025-09-05.
- NATO. 2021b. Press Conference by NATO Secretary General Jens Stoltenberg Ahead of the Meetings of NATO Defence Ministers on 17 and 18 February. 15 February. <https://www.nato.int/en/news-and-events/events/transcripts/2021/02/15/press-conference>. Accessed: 2025-09-06.

NATO. 2022. NATO 2022 Strategic Concept. Madrid, 29 June. <https://www.nato.int/content/dam/nato/webready/documents/publications-and-reports/strategic-concepts/2022/290622-strategic-concept.pdf>. Accessed: 2025-09-04.

NATO. 2022b. NATO's Data and Artificial Intelligence Review Board: Summary of the Establishment of the Board. 13 October. https://www.nato.int/cps/en/natohq/official_texts_208374.htm. Accessed: 2025-10-10.

NATO. 2023. The Secretary General's Annual Report 2022. Brussels, 21 March. <https://www.ulib.sk/files/english/nato-library/collections/monographs/secretary-general-annual-report/sgar22-en.pdf>. Accessed: 2025-11-15.

NATO. 2024. Summary of NATO's Revised Artificial Intelligence (AI) Strategy. 10 July. https://www.nato.int/cps/en/natohq/official_texts_227237.htm. Accessed: 2025-09-15.

NATO. 2025. Defence Innovation Accelerator for the North Atlantic (DIANA). https://www.nato.int/cps/en/natohq/topics_216199.htm. Accessed: 2025-09-15.

NATO Advisory Group. 2021. Annual Report 2020. March.

NATO Innovation Fund (NIF). 2025. NATO Innovation Fund. <https://www.nif.fund/>. Accessed: 2025-09-15.

Onderco, M. 2025. Navigating the AI Frontier: Insights from the Ukraine Conflict for NATO's Governance Role in Military AI. *Journal of Strategic Studies* 48(3): 602–626. doi.org/10.1080/01402390.2025.2463451.

Sayler, K. M. 2020. Artificial Intelligence and National Security. Congressional Research Service Report R45178. Washington, D.C.: Congressional Research Service.

Schmid, J.; Brummer, M.; and Taylor, M. Z. 2017. Innovation and Alliances. *Review of Policy Research* 34(5): 588–616. doi.org/10.1111/ropr.12244.

Shivakumar, S.; Wessner, C.; and Howell, T. 2025. The Limits of Chip Export Controls in Meeting the China Challenge. Center for Strategic and International Studies. <https://www.csis.org/analysis/limits-chip-export-controls-meeting-china-challenge>. Accessed: 2025-10-15.

Soare, S. R. 2021. Innovation as Adaptation: NATO and Emerging Technologies. German Marshall Fund of the United States. 4 August. <https://www.gmfus.org/news/innovation-adaptation-nato-and-emerging-technologies>. Accessed: 2025-10-23.

Stanley-Lockman, Z. 2021. Responsible and Ethical Military AI: Allies and Allied Perspectives. Washington, DC: Center for Security and Emerging Technology. <https://cset.georgetown.edu/publication/responsible-and-ethical-military-ai/>. Accessed: 2025-09-05.

Stanley-Lockman, Z., and Trabucco, L. 2024. NATO's Role in Responsible AI Governance in Military Affairs. In *The Oxford Handbook of AI Governance*, edited by J. B. Bullock, 1015–1042. Oxford: Oxford University Press.

Taddeo, M., McNeish, D., Blanchard, A., and Edgar, E. 2023. Ethical Principles for Artificial Intelligence in the Defence Domain. In *Artificial Intelligence and International Conflict in Cyberspace*, edited by F. Cristiano, D. Broeders, F. Delerue, F. Douzet, and A. G ry, 159–185. London: Routledge.

Trabucco, L., and Maas, M. M. 2023. Technology Ties: The Rise and Roles of Military AI Strategic Partnerships. SSRN preprint. SSRN: <https://ssrn.com/abstract=4629283>.