AI, Automated Systems, and Resort-to-Force Decision Making – Policy Recommendations

- **Report prepared by Professor Toni Erskine (Australian National University), Chief Investigator, *Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making Research Project,* funded by the Australian Government through a grant from the Department of Defence. Contact:** toni.erskine@anu.edu.au.[1]

Researchers: *Project CI and Workshop Co-convenor, Toni Erskine (Australian National University/Australia); Workshop Co-convenor, Steven E. Miller (Harvard University/US); Zena Assaad (Australian National University/Australia); Bianca Baggiarini (Deakin University/Australia); Maurice Chiodo (University of Cambridge/UK); Jenny L. Davis (Vanderbilt University/US); Ashley Deeks (University Virginia Law School/US); Miah Hammond-Errey (Deakin University/Australia);Yee-Kuang Heng (The University of Tokyo/Japan); Marcus Holmes (William & Mary/US); Sarah Logan (Australian National University/Australia); Paul Lushenko (U.S. Army War College/US); Dennis Müller (University of Cologne/Germany); Osonde Osoba (LinkedIn/US), Neil Renic (University of Copenhagen/Denmark); Mick Ryan (Lowy Institute/Australia); Mitja Sienknecht (European University Viadrina/Germany); Karina Vold (University of Toronto/Canada); Nicholas Wheeler (University of Birmingham/UK); Elizabeth T. Williams (Australian National University/Australia); Benjamin Zala (Monash University/Australia); Luba Zatsepina (Liverpool John Moores University/UK)*.

Introduction

The recommendations in this Policy Report arise from a two-and-half-year research project (2022-2025), entitled ***Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making***, led by Professor Toni Erskine (Australian National University) and funded by the Australian Government through a grant by the Department of Defence.

This research project has brought together **leading scholars and practitioners** working on different aspects of international politics and security, strategic and defence studies, and artificial intelligence (AI) to contribute to a multi-disciplinary study and set of **policy recommendations on the risks and opportunities of introducing AI**, machine learning (ML), and automated systems into state-level decision making on the **resort to force**. Project participants developed their findings over two international research workshops (June 2023 and July 2024) at the Australian National University (ANU), convened by Professor Toni Erskine and Professor Steven E. Miller (Harvard).

Participants had the further opportunity to share and discuss their initial research-based policy recommendations with senior Australian Government delegates from the federal civil service as part of a one-day policy roundtable (July 2024) at the ANU. We would like to thank the following delegates who participated in the roundtable and commented on the policy briefs – while noting that responsibility for their final content lies solely with the authors:

- **Captain Adam Allica** – Director, Warfare Innovation, Navy (Department of Defence)
- **Adam McCarthy** – Chief Counsel, First Assistant Secretary, Legal Division (Department of Foreign Affairs and Trade)

[1] With sincere thanks to Dr Mitja Sienknecht, Tuukka Kaikkonen, and Emily Hitchman for their assistance.

- **Ed Louis** – First Assistant Secretary, Defence Digital Group (Department of Defence)
- **Air Commodore Jarrod Pendlebury** – Director-General of Strategic Policy Futures, Risk and Outreach (Department of Defence)
- **Anthony Murfett** – Head of Division, Technology and Digital (Department of Industry, Science and Resources)
- **Andrew Seedhouse** – Chief of Space, Intelligence, National Security and Cyber (SINC) Division, Defence Science and Technology Group (Department of Defence)
- **Emil Stojanovski** – Assistant Secretary, Defence Strategic Policy Branch (Department of Foreign Affairs and Trade)
- **Captain Alastair Walsh** – Director, Net Assessments, Strategic Policy (Department of Defence)
- **Steven Yates** – Assistant Secretary, Asia Branch, Strategic Policy Division (Department of Home Affairs).

For all the potential **benefits** of AI-driven systems – which are variously able to analyse vast quantities of data, make recommendations and predictions by uncovering patterns in data that human decision makers cannot perceive, and respond to potential attacks with a speed and efficiency that we could not hope to match – challenges abound. Through this collaborative project, we have sought to address **four thematic 'complications'** that we propose will accompany the gradual infiltration of AI-enabled systems in **decisions to wage war:**[2]

- **Complication 1** relates to the displacement of human judgement in AI-driven resort-to-force decision making and possible implications for deterrence theory and the unintended escalation of conflict.

- **Complication 2** highlights detrimental consequences of automation bias, or the tendency to accept without question computer-generated outputs – a tendency that can make human decision makers less likely to use (and maintain) their own expertise and judgement.

- **Complication 3** confronts algorithmic opacity and its potential effects on the democratic and international legitimacy of resort-to-force decisions.

- **Complication 4** addresses the likelihood of AI-enabled systems impacting organisational structures and chains of command, whether degrading or enhancing strategic and operational decision-making processes.

These proposed **complications** are explored by the contributors to this project in the context of either **automated self-defence** or the use of **AI-driven decision-support systems** that inform human **resort-to force** deliberations. Each researcher has sought to identify a risk or opportunity of using AI-enabled systems in these contexts, asking how the risk can be mitigated or the opportunity promoted.

Significantly, our collective attempt to grasp the potential hazards and benefits of employing AI-driven systems to contribute to the decision to wage war draws on a **range of disciplines**. Our interventions are variously made from the perspectives of political science, international relations (IR), law, computer science, philosophy, sociology, psychology, engineering, and mathematics. We believe that this degree of interdisciplinary collaboration has produced particularly rich, productive, and challenging engagements.

[2] For an account of these 'four complications', please see T. Erskine and S. E. Miller, 'AI and the Decision to Go to War: Future Risks and Opportunities', *Australian Journal of International Affairs,* Vol. 78: 2 (2024), 135-147 (pp. 139-40).

Complication 1: The Displacement of human judgement in AI-driven resort-to-force decision making and possible implications for deterrence theory and the unintended escalation of conflict

Policy Briefs:

1.1 Professor Nicholas Wheeler (University of Birmingham) and Professor Marcus Holmes (William and Mary)
1.2 Dr Benjamin Zala (Monash University)
1.3 Dr Luba Zatsepina (Liverpool John Moores University)
1.4 Professor Ashley Deeks (University of Virginia Law School)
1.5 Dr Zena Assaad (ANU) and Associate Professor Elizabeth Williams (ANU)

Policy Brief 1.1 – Potential Benefits of AI in Nuclear Crisis Decision Making

Researchers:

- Professor Nicholas Wheeler (University of Birmingham)
- Professor Marcus Holmes (William and Mary)

Research Paper: Wheeler, Nicholas, and Holmes, Marcus, 'The Role of Artificial Intelligence in Nuclear Crisis Decision Making,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making*, Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 164-174. https://doi.org/10.1080/10357718.2024.2333814

Main Argument:

Our research explores the nuanced interplay between artificial intelligence (AI) and human decision-making in the high-stakes arena of nuclear crisis management. We argue that AI, despite its lack of emotional intelligence and experiential learning, presents unique opportunities to enhance decision-makers' ability to navigate the complexities of nuclear crises. By juxtaposing AI's data-driven insights with human intelligence's depth in emotional and creative processes, we illustrate the complementary roles each can play in fostering empathy, understanding, and, ultimately, security dilemma sensibility (SDS). Through theoretical exploration and thought experiments on historical crises such as the Cuban Missile Crisis and the Able Archer incident, we demonstrate AI's potential in mitigating misperceptions and facilitating informed, empathetic responses that acknowledge the fears and intentions of adversaries.

However, we also highlight the inherent limitations and ethical considerations of over-relying on AI, stressing the irreplaceable value of human judgment and the need for a balanced approach that leverages the strengths of both AI and human intelligence. Our conclusion underscores the importance of integrating AI as a tool within a broader strategy of crisis management that prioritizes trust-building and direct communication among decision-makers to navigate the delicate dynamics of international security and diplomacy effectively.

Policy Insights and Recommendations:

- **Potential for promotion of empathy and trust**: AI has the potential to be a valuable tool in nuclear crisis management, enhancing decision-making processes that can promote empathy and trust and reduce these escalatory pressures.
- **Potential enhancement of security dilemma sensibility (SDS)**: Decision makers need to exercise security dilemma sensibility (SDS) in times of crisis. Decision makers and diplomats exercise SDS when they are open to the possibility that the other side is behaving the way they

are because they are fearful and insecure, and crucially, recognize the role that their own actions may have played in this. Artificial Intelligence, with its data-driven analysis, might play a critical role in enhancing SDS during nuclear crises. By sifting through vast amounts of historical and real-time data, AI can help identify patterns and correlations that human analysts might overlook.

- **Risk of depersonalization of diplomacy**: The risk of over-reliance on AI is that it may lead to a depersonalization of diplomacy, where data-driven decisions overshadow the nuanced, human-centric approach that is essential in international relations. Trust and mutual understanding, often cultivated through face-to-face interactions, remain critical in diplomatic engagements. AI, no matter how advanced, cannot replicate the depth of human relationships and the trust they foster, which are often the key to resolving conflicts and preventing escalations.
- **Create balanced integration of AI and human judgement**: Ultimately, the successful integration of AI and human judgment in nuclear crisis management will depend on the ability to strike a balance that leverages the strengths of both. By maintaining human oversight and ethical standards, while also utilizing AI's analytical capabilities, decision-makers can navigate the complexities of international security with enhanced insight and precision, leading to more effective and sustainable conflict resolution strategies.

Policy Brief 1.2 – Risks of AI in Nuclear Command and Control

Researcher: Dr Benjamin Zala (Monash University)

Research Paper: Zala, Benjamin, 'Should AI stay or should AI go? First strike incentives & deterrence stability,' Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making, Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 154-163. https://doi.org/10.1080/10357718.2024.2328805

Main Argument:

The risks of deploying AI in nuclear command and control must be analysed against the backdrop of a larger trend in global nuclear politics – the increasing prominence of strategic non-nuclear weapons (SNNW) in issues of deterrence stability. SNNW (such as conventional precision-strike missiles, missile defence, anti-satellite weapons, etc.) make nuclear deterrence relationships more fragile and crisis signalling more complex.
Two broad sets of risks can be identified and distinguished from each other. First, automation in military deployments, or taking the human 'out of the loop' in the decision to use a nuclear weapon or SNNW. Second, risks arising from the use of AI in informing human decision-making (particularly early warning threat assessments).

There are also opportunities which can be exploited in which AI may help to restabilise deterrence relationships. E.g., AI and machine learning can be used to improve techniques used for anomaly detection through pattern recognition. Such techniques can increase a state's confidence in the survivability of its second-strike capabilities.

Policy Insights and Recommendations:

- **Apply risk assessments broadly**: Risk assessments relating to the deployment of AI and machine learning need to be applied not only to obvious areas such as nuclear launch orders, but also less obvious areas such as early warning intelligence assessments (including by non-nuclear allies) and strategic non-nuclear weaponry (SNNW) capabilities (also, including by non-nuclear allies)

- **Limit reliance on AI-assisted warning data:** The key to balancing the benefits of incorporating AI into early warning against the risks, is limiting what AI-assisted warning data is used for. Tasks such as calculating effective evasive manoeuvres in the event of an attack and using pattern recognition and anomaly detection to improve arms control verification should be prioritised in AI research.
- **Pursue informal arms control and confidence-building:** Informal arms control and confidence-building measures should be pursued relating to AI and nuclear command and control today. These include regular dialogues aimed at fostering common understandings of potential dangers and establishing red lines as well as information exchange mechanisms. Unilateral measures, such as moratoriums, should continue and be expanded.

Policy Brief 1.3 - Dangers of AI Competition in Nuclear Command and Control

Researcher: Dr Luba Zatsepina (Liverpool John Moores University)

Research Paper: Zatsepina, Luba, 'Waltzing into Uncertainty: AI in Nuclear Decision Making and the Dangers of a New Arms Race,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

The integration of AI into nuclear decision making undermines the stability assumed by rational deterrence theory. While there are potential benefits to AI supporting command and control processes, we must ensure that human decision-making remains central to determining when and how nuclear-weapon states resort to the use of their arsenals. Over-reliance on AI systems in this context could lead to miscalculations or unintended escalations in the use of force. Furthermore, the competition for AI superiority in nuclear command and control *can* and *will* lead to an arms race, as states strive to outmatch each other's technological capabilities, potentially escalating tensions. This technological competition could make decisions to use nuclear weapons (both strategic and tactical) less predictable, thereby increasing the risk of misinterpretation and miscalculation.

Policy Insights and Recommendations:

- **Incorporate human-in-the-loop safeguards**: ensure AI systems in nuclear command and control are always overseen by human operators. This would prioritise human judgement in critical decision-making stages.
- **Encourage AI risk assessments in national security protocols**: conduct regular risk assessments of AI integration in nuclear command structures, with a focus on ethical risks, system vulnerabilities, and crisis scenarios.
- **Encourage collaborative international research** on the safety, ethics, and governance of AI in military applications.
- **Implement confidence-building measures** (e.g., data exchange, joint research on AI safety) that focus on AI transparency among nuclear states.

Policy Brief 1.4 – Mitigating AI Errors in Resort-to-Force Decision Making

Researcher: Professor Ashley Deeks (University of Virginia Law School)

Research Papers:

1. Deeks, Ashley, 'Delegating War Initiation to Machines,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making*, Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 148-153. https://doi.org/10.1080/10357718.2024.2327375
2. Deeks, Ashley, 'State Responsibility for AI Mistakes in the Resort to Force,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

There are several situations in which the use of AI or autonomy in resort-to-force decisions may produce errors, including where the system is poorly trained; where another state poisons the system's data; or where two AI-driven systems interact in unintended ways. Though the law here is nascent, those errors likely will constitute an internationally wrongful act under the *jus ad bellum* unless (1) the state has made an honest and objectively reasonable mistake or (2) a circumstance precluding wrongfulness exists. To that end, the Australian Government Department of Defence might consider the following steps to advance the state of the law and minimize the risk of inadvertent conflict arising from such errors.

Policy Insights and Recommendations:

- **Clarify legal standards of care**: The DoD (and other Five Eyes allies) should consider taking public positions that would clarify the legal standards of care for acts covered by the *jus ad bellum* and international humanitarian law.
- **Adopt robust security and cyber hygiene**: The DoD should adopt robust security and cyber hygiene against AI data poisoning and hacking to ensure that it can meet a *jus ad bellum* standard of care of a good faith and objectively reasonable action.
- **Impose contractual requirements**: The DoD should consider imposing contractual requirements on contractors producing AI systems for it that ensure that the DoD has good visibility into data sets, training processes, and system limitations of the AI systems it acquires.
- **Test interoperability**: Consistent with the proposed "Five AIs Act" in the U.S. Congress, consider testing the interoperability of Five Eyes AI systems in both cooperative and adversarial postures to understand how the systems will interact with each other.
- **Seek cooperation**: Consider seeking cooperation among close allies in the field of AI testing, evaluation, validation, and verification.
- **Provide legal guidelines for use-of-force delegation to autonomous systems:** Urge senior leadership to provide clear instructions about the propriety of delegating the use of force to autonomous systems under domestic law and about the standards under which such delegation may occur.
- **Do not replace humans with machines unless they can achieve higher accuracy rates.**: Recognize that adversaries may be more likely to forgive human error than machine error, which means that states should be hesitant to deploy systems that do not produce accuracy rates higher than humans.

- **Maintain transparency regarding after-action reviews.** Commit to being transparent and deliberate about after-action reviews of any AI errors that occur in the field; consider using civilian casualty review processes as a model.

Policy Brief 1.5 – Safety, AI Decision-Support Systems, and the Resort to Force

Researchers:

- Dr Zena Assaad, School of Engineering (ANU)
- Dr Elizabeth T. Williams, School of Engineering (ANU)

Research Paper: Assaad, Zena, and Williams, Elizabeth T., 'Technology and tactics: The intersection of safety, AI, and the resort to force,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

AI adds complexity to decision-support systems (DSS) used to inform resort-to-force decision-making. This in the form of increased interactivity, nonlinearity, software complexity, and dynamics of scale (time, geography, and so on). Interactive and nonlinear complexity can result in misinformed decision making. Increased software complexity leads to a loss of fundamental knowledge in how systems operate, and dynamic complexity creates challenges with ubiquity and varying timescales. We discuss how this complexity impacts safety, trust and liability related to the use of DSS.

We remind readers that humans shape the design, training, bounds, use, and evolution of AI used in such systems – something that is often forgotten when discussing AI, due to the anthropomorphized language used to describe AI system properties and performance. We discuss how human decision-making plays a considerable role in all aspects of such systems and is in turn shaped by such systems. We argue that human decision-making be placed at the forefront of considerations around how to consider and manage safety considerations relevant to the introduction and proliferation of AI in DSS used to inform resort-to-force decision-making.

Policy Insights and Recommendations:

- **Embrace holistic approach to AI-safety**: The notion of safety of AI encompasses both technical and socio-technical considerations. When assessing the potential safety challenges of AI-enabled DSS, it should be considered holistically to include broader considerations such as security, trust and liability.
- **Adopt expert risk assessment process**: There is an intersection between the safety of AI-enabled DSS and resort-to-force decisions. Human decision making is impacted when safety is compromised. When implementing AI-enabled DSS, these safety considerations should be formally captured through a risk assessment process conducted by appropriately trained experts, so appropriate mitigations can be employed. This process should also identify points (or triggers) in the system life cycle that would necessitate a re-assessment of risks and required mitigations.
- **Identify and document the roles and responsibilities of systems and humans**: The roles and capabilities of AI-enabled tools are commonly misunderstood or embellished, particularly when determining what human roles and responsibilities are in relation to these tools. When implementing AI-enabled DSS, roles and responsibilities of both the system and the humans operating alongside the system should be clearly identified, documented, and understood by those responsible for making use of such systems.

- **Implement policy that supports a safety culture**: Given the magnitude of the risks in question, States should consider ways of implementing policy that supports a safety culture for people and organisations responsible for systems making use of AI in such high-risk applications. This may include setting up an independent regulator responsible for ensuring the risk assessments recommended above are conducted appropriately and ensure that any risk mitigation strategies are properly applied. Approaches used to control other high-hazard technologies (e.g. nuclear) may be useful as a starting point for discussing how to achieve this.

Complication 2: The consequences of automation bias

Policy Briefs:

2.1 Lieutenant Colonel/Dr Paul Lushenko (US Army War College)
2.2 Professor Jenny L. Davies (Vanderbilt University)
2.3 Dr Neil Renic (University of Copenhagen)
2.4 Professor Karina Vold (University of Toronto)
2.5 Professor Toni Erskine (ANU)

Policy Brief 2.1 – Enhancing Trust in AI Used in the War Room

Researcher: Lieutenant Colonel/Dr Paul Lushenko (Cornell University and US Army War College)

Research Paper: Lushenko, Paul, 'AI, Trust, and the War-Room: Evidence from a Conjoint Experiment in the US Military,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

The purpose of this research was to explain what shapes military trust for AI during crisis escalation. Despite predictions that AI is 'game-changing', it is unclear what shapes military trust in AI during strategic-level deliberations. To study this question, I surveyed an elite sample of officers attending the US Army and Naval War Colleges, assessing how variation in different features of AI shape their attitudes of trust. My results suggest that trust is a function of a tightly calibrated set of considerations relating to how AI is used, for what outcomes, and with what oversight.

I found that soldiers generally trust AI and that their attitudes are moderated by (1) several technical specifications, including non-lethal use, maximum precision, and human oversight; (2) perceived effectiveness measured in terms of equitable protection for civilians and soldiers, as well as contribution to mission success; and, (3) regulatory oversight, particularly levied internationally.

Policy Insights and Recommendations:

- **Temper expectations** for AI in the 'war-room'. While AI may be shifting the character of war, or how it is fought, it is not shifting the nature of war, or why it is fought.
- **Consider the multidimensionality of trust**: Recognize that soldiers' trust in AI is not a forgone conclusion. Rather, it is complex and multidimensional, and further complicated by biases, uncertainty, and lack of education.
- **Promote research on trust in AI**: In terms of research, military attitudes of trust in AI needs more cross-national investigation.

- **Modernize AI policy for trust**: In terms of military modernization, align warfighting concepts, doctrine, and regulations and policies that govern AI to reflect soldiers' attitudes, which promises to engender more trust.
- **Increase AI literacy of military officers**: In terms of professional military education, require the Australian War College to ensure officers have an understanding of data, data analytics, and AI, including decision-support algorithms.
- **Interrogate norm compliance**: In terms of governance, explain how policies on increasingly autonomous capabilities coincide or diverge from international norms and laws informing their use.

Policy Brief 2.2 – Experts-in-the-Loop and Resort-to-Force Decision Making

Researcher: Professor Jenny L. Davis, Professor of Sociology (Vanderbilt University); Honorary Professor of Sociology (ANU)

Research Paper: Davis, Jenny L. (2024). 'Elevating humanism in high-stakes automation: experts-in-the-loop and resort-to-force decision making,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making,* Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 200–209. https://doi.org/10.1080/10357718.2024.2328293

Main Argument:

When AI are used for high-stakes decisions, like the resort-to-force, humans with domain expertise are vital. 'Experts-in-the-loop' refers to an organisational structure that emphasises human expertise when AI are involved in decisions. AI escalate the volume of information, speed of action, and scale of effects.

Domain expertise is necessary to interpret that information, deliberate on action, and meaningfully consider possible outcomes. The alternatives to experts-in-the-loop are full automation (relying on AI to make and execute decisions) and technicians-in-the-loop (personnel with primarily technical skillsets). Both options are inadequate for resort-to-force decision making.

Policy Insights and Recommendations:

- **Embed experts in decision structures**: Enshrine an expert-in-the-loop organisational structure—i.e., high-level experts as core decision makers (e.g., high ranking officers and intelligence specialists).
- **Prohibit automation** of resort-to-force decisions.
- **Increase AI literacy of domain experts**: Provide and require basic technical training for high-level domain experts so they understand the logics of AI and can thus incorporate AI decision inputs from an informed position.
- **Provide training for domain experts**: Sustain substantive training for, and assessment of, high-level experts to bolster and ensure substantive competencies.

Policy Brief 2.3 – Moral and Political Wisdom in AI-assisted Decision Making

Researcher: Dr Neil Renic, Centre for Military Studies (University of Copenhagen)

Research Papers:

1. Renic, Neil, 'Tragic Reflection, Political Wisdom, and the Future of Algorithmic War,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making*, Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 247-256. https://doi.org/10.1080/10357718.2024.2328299
2. Renic, Neil, 'AI Optimized Violence and the Suffocation of Moral and Political Wisdom,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main argument:

The speed, inflexibility, and false confidence of algorithmically assisted decision making cultivates an insensitivity to the tragic qualities of violence. This dulling of the tragic sensibility imperils not only human knowledge and skill, but also wisdom, and is likely to lead to more imprudent and immoral uses of force, not less.

Policy Insights and Recommendations:

- **Prioritize resilience.** Underpinning much of the optimism over AI and machine learning in resort-to-force decision making is an exaggerated faith in optimization. Through optimization, proponents of AI hope to identify and eliminate inefficiencies and streamline decision-making processes–not unimportant when the merits and risks of violence have to be calculated in time-sensitive circumstances. As we learned during the COVID-19 crisis, however, when our "just enough, just in time" global supply chains came undone, optimization is intrinsically brittle. If we over-optimize our resort-to-force decision-making, these systems will be vulnerable when conditions shift in unpredictable ways (as they predictably will). For systems to be resilient in the face of change, some degree of "slack" must be maintained. This will mean inefficiency, but of the meaningful sort, allowing human agents to innovate when faced with novel challenges that confound our algorithmic tools.
- **Test for moral and political suffocation points.** Testing resort-to-force AI systems against all possible scenarios which may arise after deployment is likely impossible. More attention can and must be paid, however, to identifying the points at which meaningful human control can no longer be exercised within the decision chain. "Control" should be understood broadly in this context, to include the inclination to ethically reason and the capacity to ethically intervene to override AI systems where necessary. Technology, including virtual reality exercises, can play an important role in revealing these limits. Virtual reality affords us the freedom to experiment with speed to observe the effects of greater and lesser degrees of compression on the quality of resort-to-force decision-making. The same can be done with the complex digital environments these decision makers will inhabit. Explicit study is needed to clarify which human-machine interfaces best preserve the agency of users and which habituate problematic patterns of action.

Policy Brief 2.4 – Strategic Decision-Making Advantages of Non-Autonomous AIs

Researcher: Professor Karina Vold (University of Toronto)

Research Paper: Vold, Karina, 'Human-AI Cognitive Teaming: using AI to support state-level decision making on the resort to force,' Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making, Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 229-236. https://doi.org/10.1080/10357718.2024.2327383

Main Argument:

Many AI systems are non-autonomous systems rather than fully autonomous agents. Non-autonomous AI systems make use of sophisticated machine learning techniques but only semi-automate a task, e.g., language translators; prompt-enabled generative systems, e.g., ChatGPT, Dall-E.

Many non-autonomous AIs are built to help humans complete cognitive tasks and aid our cognitive capacities, e.g., memory, attention and search, planning, communication, comprehension, emotion and self-control, navigation, conceptualization, quantitative and logical reasoning, etc.
Non-autonomous AIs are often overlooked, but can provide critical strategic advantages to decision makers who deploy them successfully.

Policy Insights and Recommendations:

- **Regulate non-autonomous AI**: While autonomous AI agents, e.g., lethal autonomous weapons systems (LAWS), need regulation, so do non-autonomous AI systems, which leave humans vulnerable to new forms of influence, moral and cognitive atrophy, and undermined responsibility.
- **Monitor non-autonomous AI**: Regulation should include continual monitoring as non-autonomous systems can become integrated parts of their human user's overall cognitive system, hence changes to how the 'tool' functions or if it ceases to function entirely can have critical impacts on the user's cognitive functions.
- **Recognise risks of non-autonomous AI**: Non-autonomous AI systems that model and monitor human behaviour to find targeted interventions that optimize some metric of cognitive performance can also denigrate into forms of surveillance and manipulation.

Policy Brief 2.5 – AI-Driven Decision-Support Systems and Norms of Restraint

Researcher: Professor Toni Erskine, Coral Bell School of Asia Pacific Affairs (ANU)

Research Papers:

1. Erskine, Toni, '*Before* Algorithmic Armageddon: Anticipating Immediate Risks to Restraint when AI Infiltrates Decisions to Wage War,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making,* Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 175-190. https://doi.org/10.1080/10357718.2024.2345636 .
2. Erskine, Toni, 'AI and the Future of IR: Disentangling Flesh-and-Blood, Institutional, and Synthetic Moral Agency in World Politics', *Review of International Studies*, 50: 3 (2024), pp. 534-559 (esp. pp. 553-55, 556-58).

Main argument:

AI-enabled systems will steadily infiltrate resort-to-force decision making. This will likely include decision-support systems (DSS) recruited to assist with crucial deliberations over the permissibility of

waging war. Potential benefits abound in terms of enhancing individual and institutional capacities for cognition, analysis, and foresight. Yet, we have reason to worry. Our interaction with these systems – as citizens, political and military leaders, states, and formal organisation of states – would also court significant risks. Specifically, reliance on DSS that employ machine-learning techniques would threaten to undermine our adherence to international norms of restraint in two distinct ways: (i) by creating the reassuring illusion that these AI-driven tools are able to replace us as responsible agents (the 'risk of misplaced responsibility'); and (ii) by inserting unwarranted certainty and singularity into complex *jus ad bellum* judgements (the 'risk of predicted permissibility'). If unaddressed, each proposed risk would make the initiation of war appear more permissible in particular cases and, collaterally, contribute to the erosion of hard-won international norms of restraint.

Policy Insights and Recommendations:

- **Educate decision makers about the nature and limitations of AI-driven DSS.** Political and military leaders whose deliberations over the legitimacy of waging war would be influenced by the predictions and recommendations of AI-enabled decision-support systems must be educated about how these systems function (through statistical inference), their corresponding limitations (they lack capacities for understanding, self-reflection, and judgment), and the status of their outputs (as data-driven guesses). This would begin to guard against the dual tendencies to defer to the outputs of DSS and to disregard alternative possibilities.

- **Reinforce where responsibility lies for adhering to norms of restraint.** It is necessary to reiterate and reinforce that responsibility for resort-to-force decisions remains with the state's political and military leaders and relevant executive and legislative bodies. Guidance from AI-driven DSS neither redirects nor dilutes these existing loci of responsibility. This needs to be stated explicitly.

- **Design AI-driven DSS to promote a more accurate perception of their capacities.** DSS must be designed so that they cannot easily be mistaken for responsible agents in themselves. This will involve, for example, refraining from anthropomorphising them, building in warnings that remind users of their limitations, and incorporating cues that reinforce human agency and responsibility.

- **Establish 'supplementary responsibilities of restraint'**. We have robust international norms of restraint in terms of strict legal and moral guidelines on when states can legitimately engage in armed conflict. These international norms need to be updated and elaborated to encompass additional imperatives regarding *whether, when,* and *how* to employ AI-driven DSS in judgements over war initiation.

Complication 3: Algorithmic opacity and its potential effects on democratic and international legitimacy

Policy Briefs:

3.1 Dr Bianca Baggiarini (Deakin University)
3.2 Dr Sarah Logan (ANU)
3.3 Dr Osonde Osoba (RAND Corporation)
3.4 Dr Miah Hammond-Errey (Strat Futures and Deakin University)

Policy Brief 3.1 – Democratic Legitimacy in AI-enabled Resort-to-Force Decisions

Researcher: Dr Bianca Baggiarini (Deakin University)

Research Papers:

1. Baggiarini, Bianca, 'Algorithmic war and the dangers of in-visibility, anonymity, and fragmentation,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making,* Special Issue of *Australian Journal of International Affairs*, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 257-265. https://doi.org/10.1080/10357718.2024.2333824
2. Baggiarini, Bianca, 'A king above the law: 'autocratic intelligence,' resort-to-force decision making, and democratic legitimacy in war,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

Increasing militarization, authoritarian politics, and democratic decline together present urgent challenges for policymakers and security/defence practitioners. Meanwhile, militaries are embracing AI-enabled technologies, many of which are contributing to, or stem from, the challenges cited above, and which transform the nature of decision making in war. These challenges are socio-technical in nature; the technological aspects cannot be neatly separated from the social and political justification and effects. Within liberal democracies, resort-to-force decision making is expected to be legitimate. Yet, AI transforms how democratic legitimacy in war is perceived and practiced.

While transparency is a part of communicating and perceiving democratic legitimacy in war, technological understandings of transparency should not be prioritized, or come at the cost of, sociopolitical forms of transparency. Since war and democracy are opposed, tremendous effort is required to ensure that authoritarian ideals (which may be enabled by AI) do not influence resort to force decision making.

Policy Insights and Recommendations:

- **Establish social and political transparency**: To increase the perception of democratic legitimacy in resort-to-force decision making, social and political forms of transparency – to the extent possible – should accompany any/all attempts to render AI-enabled technology transparent.
- **Strengthen commitment to international law**: Likewise, the Australian Defence Force (ADF) should actively reassess and strengthen its commitment to its people, as well as its ethical and legal obligations under international law, as it simultaneously integrates AI-enabled platforms.
- **Enhance (international) cooperation**: Increased militarization is a threat to democratic legitimacy and should be countered through enhanced diplomatic communications, international partnerships aimed at peace and stability, and other non-militarized activities.

Policy Brief 3.2 – LLMs in Intelligence Analysis

Researcher: Dr Sarah Logan (ANU)

Research Paper: Logan, Sarah, 'Tell me what you don't know: large language models and the pathologies of intelligence analysis,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making,* Special Issue of *Australian Journal of International Affairs*, guest

edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 220-228. https://doi.org/10.1080/10357718.2024.2331733

Main Argument:

This paper argues that large language models (LLMs) intensify two existing pathologies of intelligence analysis: information scarcity and epistemic scarcity. On the former, it argues that data collection for the purposes of building LLMs is dominated by commercial actors who have few commercial incentives to cooperate with intelligence agencies. This means that intelligence agencies cannot easily acquire the data to build their own LLMS and cannot easily interrogate the data they acquire from vendors to assess its credibility or vulnerability to outside interference. On the latter, the paper argues that even if intelligence agencies were able to secure training data to build their own LLMs, much of this training data is wrought from online sources only, meaning it is dominated by English-language text which, within the inherent limits of AI, limits its ability to deliver truthful analytic judgments which are capable of informing decisions to go to war in an ethical and accountable manner. The paper ends by arguing that Western states such as Australia are at an operational disadvantage compared to authoritarian states, which can harvest online data with greater legal compliance from private actors.

Policy Insights and Recommendations:

- **Limit epistemic pathologies of LLMs**: Clearly determine Australian defence and intelligence policy towards either a) procurement of or b) state development of LLMs, or c) a combination of both and use this guidance to develop policy which seeks to limit the epistemic pathologies of LLMs in autonomous decision-making.
- **Commit to procurement guidelines and oversight**: Commit to sector-wide procurement guidelines and oversight of generative AI tools used in decision-making chains.
- **Regulate data markets**: Commit to regulating data markets and Australian access to those markets through alliance relationships.
- **Bridge AI safety and national security**: Commit to engaging the national security sector in ongoing Australian WoG discussions on AI safety rather than carving out wholesale exceptions for that sector. For example, ongoing discussions via the Department of Industry, Science and Resources offer important perspectives to national security decisionmakers, even if the final set of AI guidelines is not applied to the national security sector.
- **Encourage sector-wide discussion on the accuracy of AI tools**: The recent IGIS report on agency use of artificial intelligence notes that many agencies are proactive in organising AI governance boards and are attentive to questions of the ethical use of AI. I would argue strongly that ethical use of AI in the national security context includes well-informed, sceptical use. IGIS notes that many agencies are aware of the risks of bias in training datasets and seek to incorporate AI use into appropriate legal and ethical frameworks in attempt to mitigate bias. However, if they are not already occurring, sector-wide discussions about the accuracy of AI tools including but not limited to ethical problems of bias should be encouraged. These should include discussions of the broader epistemic limitations of language and sourcing and perhaps include a sandbox-style testing and demonstration ground to inform procurement decisions and use guidelines.

Policy Brief 3.3 – Responsible AI Governance and Military Decision Making

Researcher: Dr Osonde Osoba (RAND Corporation)

Research Papers:

1. Osoba, Osonde, 'A Complex-Systems View on Military Decision Making,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making*, Special Issue of

Australian Journal of International Affairs, guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 237-246. https://doi.org/10.1080/10357718.2024.2333817

2. Osoba, Osonde, 'Responsible AI Governance for Military Decision Making: A Proposal for Managing Complexity,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

Military decision-making institutions are complex adaptive systems undergoing massive shock as they adapt to the widespread integration of AI and automation tools. We need effective frameworks, both to understand and better manage the effects of this shock. The use of analogies to other complex adaptive decision systems can give more perspective on common AI concerns like deskilling and algorithmic opacity. Insights from this line of thinking suggest three points: 1) human-machine teams possess a form of cognitive diversity that can be leveraged for more efficient decision-making or exploited to poison information flows; 2) deskilling is the flipside of beneficial specialization. Specialization on operational tasks in human-machine teams may improve decision-making performance; 3) technical explanations for algorithmic opacity will **not** solve accountability concerns.

Finally, I explore the value of adopting responsible AI (RAI) governance programs to help manage the complexity induced by AI integration. I argue that even parsimoniously-specified RAI governance programs may have value in fostering a culture of accountability in the use of AI to support military decision-making. However, I raise a hypothetical scenario in which inefficient RAI processes undermine deterrence calculi by lengthening a deterring state's decision timeline.

Policy Insights and Recommendations:

- **Fund Research and Development**: Invest in R&D to maximize the benefits of human-machine cognitive diversity.
- **Implement Responsible AI**: Implement RAI governance programs that carefully balance accountability with operational efficiency.
- **Include safety mechanisms**: Perform regular red-team exercises to ensure that the integration of AI in decision-making institutions does not induce systemic blind spots and vulnerabilities in military decision-making.

Policy Brief 3.4 – The Impact of the Tech Stack on Decisions to Go to War

Researcher: Dr Miah Hammond-Errey (Strat Futures and Deakin University)

Research Paper: Hammond-Errey, Miah, 'Architectures of AI: Tech power broking war?,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

This paper examines big data and AI as a broader ecosystem, across the 'stack', defining the components of the tech stack and their underlying infrastructure (data, connectivity, compute capacity and workforce) as the 'architecture of AI.' It highlights the role that architectures of AI play in deliberations surrounding the resort to force and their potential to change the calculus (political and otherwise) of going to war. To do so, this research examines the concentration of power, diffused national security decision making, and AI in our social media and news information environment.

Policy Insights and Recommendations:

- **Understand the AI tech stack and its fragilities**: Increase understanding of the tech stack, the inherent interdependencies and vulnerabilities in tech ecosystem, and the fragility of the architecture of AI systems, through
    - **Technology literacy training programs** designed specifically for politicians as well as policy, intelligence, and military leaders.
    - **Research** to map the architecture of Australian digital infrastructure and AI capabilities.
    - **Research** to develop a comprehensive picture of the architecture – physical and digital – that underpins AI, including critical dependencies and vulnerabilities for Australia and our region and how access and power are distributed.
- **Fund forecasting of future technology dependencies** (for Australian government, ADF and NIC functions).
- **Fund research on social media** and its impact on functions of government, including democracy, its ability to influence public decisions on the resort to force, and foreign interference.
- **Recognize AI's role in war**: Significantly increase awareness of government reliance on the architecture of AI, especially for critical government functions and functions of war.
    - **Educate** through research, media outreach, and training.
    - **Increase intelligence collection** on critical capabilities.
- **Increase national investment** to build public sovereign capabilities, where needed.
- **Increase technology literacy** across government. Increase the depth and scope of understanding of the tech ecosystem, tech policy, and the impacts of technology on governing, security, warfare, and public safety across ADF, NIC, and APS.
- **Increase technology policy awareness**: Increase awareness of the role of technology policy in Australia–and increasingly global technology policy as well as technology multi-lateral forums–in affecting AI capabilities and dependencies across the whole of government.

Complication 4: AI-enabled systems and their impact on organisational structures

Policy Briefs:

4.1 Major General Mick Ryan (Lowy Institute)
4.2 Dr Maurice Chiodo (University of Cambridge), Dennis Müller (University of Cologne) and Dr Mitja Sienknecht (European New School of Digital Studies/European University Viadrina)
4.3 Dr Mitja Sienknecht (European New School of Digital Studies/European University Viadrina)
4.4 Professor Yee-Kuang Heng (University of Tokyo)
4.5 Professor Toni Erskine (ANU) and Professor Jenny L. Davis (Vanderbilt University)

Policy Brief 4.1 – Improving Adaptive Culture and Wartime Decisions through AI

Researcher: Major General Mick Ryan, Senior Fellow for Military Studies (Lowy Institute)

Research Paper: Ryan, Mick, 'The Fifth Element: Algorithmic Support to Adaption Before and During War,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

The capacity for rapid learning and evolution is a human capability which must be applied now in a military profession and in military institutions that are saturated with new technologies which are also quickly changing. Every wartime decision can and should be informed by previous decisions, and thus, improved through effective adaptive cultures. This might be improved further through AI decision-support tools.

Key wartime decisions that might be improved by AI-enabled adaptation include decisions about ethical use of force, balancing tactical and strategic forces, achieving optimal force structures of crewed and uncrewed systems, prioritising munitions, equipment and personnel as well as training and education. But learning and adaptation is not just a wartime concern. It is the processes, technologies, leadership philosophies and cultures put in place between wars that provide the foundation for military adaptation in war. Decision-making on peacetime functions, such as readiness, testing different force structures and equipment procurements, training and education, logistics and personnel management and the strategic management of alliance interactions might also be improved through better adaptive processes that employ AI.

This paper proposes an evolved concept for multi-level, individual and institutional military adaptation, through fusion of new learning processes and Artificial Intelligence (AI) to speed up and enhance the quality of military adaptation and strategic decision-making. This transformation of the learning cultures and processes in military institutions has very little to do with technology, however. The larger and most important role is played by humans. The success of enhanced adaptation through AIU support will be almost entirely driven by human decision-making, processes and culture.

Policy Insights and Recommendations:

- **Set (and evolve) measures of effectiveness.** If AI-enabled adaptive capacity is to work effectively, measures of military effectiveness must guide which direction adaptation might take. These need to be developed at the tactical, operational and strategic levels to guide development and implementation of AI-enabled adaptation.
- **Know where adaptation relevant data is found, stored and shared.** An enhanced adaptive stance in military institutions must have enhanced data awareness as a foundation. Institutional measures will be an important element, but so too will data discipline in tactical units and by individuals. As such, data awareness and management will need to become one of the basic disciplines taught to military personnel.
- **Explicitly embrace adaptation.** Senior institutional leaders must nurture people and formations that are actively learning and capable of changing where it is safe and effective to do so. This culture must begin with clear statements about the leadership environment, and its tolerance for risk and new ideas.
- **Scale AI support from individual to institution.** There is unlikely to be a one size fits all algorithm or process that can enhance learning and adaptation at every level of military endeavours. A virtual arms room of adaptation support algorithms will be necessary in an institution-wide approach to adaptation.
- **Connect tactical learning with strategic learning.** The observation and absorption of lessons needs to be part of normal military interaction rather than a separate and parallel ecosystem

that often has difficulty inserting itself into strategic decision making. Tactical learning must be connected with strategic learning. Human processes and committees must evolve to improve this interaction.

Policy Brief 4.2 – Educating Actors for the Responsible Military Use of AI

Researchers:

- Dr Maurice Chiodo, Centre for the Study of Existential Risk (University of Cambridge)
- Dennis Müller, Centre for the Study of Existential Risk (University of Cologne)
- Dr Mitja Sienknecht (European New School of Digital Studies/European University Viadrina)

Research Papers:

1. Müller, D., Chiodo, M., & Sienknecht, M. (2024). Integrators at War: Mediating in AI-assisted Resort-to-Force Decisions. Presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia. *Preprint at SSRN-id5045155, https://dx.doi.org/10.2139/ssrn.5094539*
2. Chiodo, M., Müller, D., & Sienknecht, M. (2024). Educating AI developers to prevent harmful path dependency in AI resort-to-force decision making. *Australian Journal of International Affairs*, *78*(2), 210-219. https://doi.org/10.1080/10357718.2024.2327366
3. Chiodo, M., & Müller, D. (2023). Manifesto for the Responsible Development of Mathematical Works - A Tool for Practitioners and for Management. *Preprint at arXiv:2306.09131v1.* https://doi.org/10.48550/arXiv.2306.09131

Main Argument:

There are three main groups involved in AI-assisted decision making: ***Developers***, who develop the AI-based tool. ***Integrators***, who implement the technology into the organisational and functional structure of the system. And ***Users***, who use the AI in a given situation. Each of these groups requires education in the responsible and safe use of AI, along the lines of the "10 Pillars of responsible development of mathematical works". Integrators face a particular "sandwich position" between developers and users and need to understand and have expertise in both domains. However, the role, and even existence, of integrators is often overlooked, invisible, or completely unknown.

Policy Insights and Recommendations:

- **Educate actor groups:** Educate each relevant group in the AI lifecycle (developers, integrators, users) along the '10 Pillars of responsible development of mathematical work' (see Table 1 from [1] for an overview, or [3] for full details).
- **Conduct research on power imbalances** between the three actor groups.
- **Provide recommendations on how to hire integrators.**
- **Fund and support research** on the integration of AI in strategic decision-making.
- **Provide (minimum) standards** for the responsibilities of AI developers and integrators.
- **Implement requirements for facilitating discussions** between developers, integrators and users during the development process, integration process, and longer-term maintenance process.

- **Provide well-defined accountability guidelines** and rules for who is accountable if something goes "wrong".

Policy Brief 4.3 – AI, Resort-to-Force Decision Making, and Responsibility Gaps

Researcher: Dr Mitja Sienknecht (European New School of Digital Studies/European University Viadrina)

Research Papers:

1. Sienknecht, Mitja, 'Proxy Responsibility: Addressing Responsibility Gaps in Human-Machine Decision Making on the Resort to Force,' *Anticipating the Future of War: AI, Automated Systems, and Resort-to-force Decision Making*, Special Issue of *Australian Journal of International Affairs,* guest edited by Toni Erskine and Steven E. Miller, Vol 78, No 2 (2024): 191-199. https://doi.org/10.1080/10357718.2024.2327384
2. Sienknecht, Mitja, 'Proxy Responsibility for AI-based Decisions in the Resort-to-Force,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

One of the key challenges in integrating AI-based technologies into decision-making processes is the emergence of a potential responsibility gap – that is, the absence of a clearly responsible human agent for decisions made by machines.

To address this issue, I propose the concept of proxy responsibility. Proxy responsibility is relevant in situations where the direct causal agent of an action—the machine—cannot qualify as a morally or legally responsible subject. In such cases, "second-order" or indirect responsibility must be established. To this end, the integration of an AI department at the institutional level is advisable, serving in an advisory capacity and bringing together comprehensive expertise. Such an institutional response would ensure that moral responsibility can still be meaningfully assigned to humans for decisions made or influenced by AI systems.

Policy Insights and Recommendations:

- **Prohibit full autonomous weapons systems**: Fully autonomous weapons systems without a human in the loop are ethically untenable and should be banned internationally.
- **Implement an advisory body**: One institutional way to establish proxy responsibility is to establish a state-level AI department at the nexus of the political, military, legal, and economic systems that integrates technical, political, and ethical competence and expertise, and advises the respective groups in the decision-making process on the resort to force.
- **Horizontally integrate state-level AI department**: This department should be horizontally integrated into the organisational structure of the military, with strong links to the political, legal, and economic systems.
- **Establish democratic oversight**: Such a department would serve as a mechanism for democratic oversight, providing ethical guidance on such dynamic and far-reaching decisions as the one to resort to force. It should be composed of experts in civil society, law, ethics, technology, integrators, decision-makers, and developers.

Policy Brief 4.4 – Upskilling Human Analysts through Education

Researcher: Professor Yee-Kuang Heng, Graduate School of Public Policy (University of Tokyo)

Research Papers:

1. Heng, Yee-Kuang, 'Upskilling human actors against AI automation bias in the resort-to-force: Education, Challenge, and Institutional functions,' presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.
2. Heng, Yee-Kuang, 'Building futures literacy: Nudging civil servants to cope with uncertainties and threats,' *European Journal of International Security,* Published online (2024): 1-18. https://doi:10.1017/eis.2024.40

Main Argument:

Upgrading AI literacy and "upskilling" human analysts through education must not be overshadowed by relentless attempts to better train AI models. Human-machine teams can constantly challenge AI-enabled groupthink. Mindsets and institutional structures in government can be tweaked to reinforce education and challenge functions.

Policy Insights and Recommendations:

- **Inform AI literacy through future literacy**: In-house and outsourced capacity-building in "futures literacy" can inform "AI literacy" programmes tailored for human-machine intelligence analysis teams.
- **Embed a challenge function**: Australia's net assessment capabilities after the 2023 Defence Strategic Review (DSR) should embed a "challenge function" focused on AI-enabled intelligence analysis and human-machine interface.
- **Institutionalise monitoring**: Build mindsets, protocols, institutional cultures, and inter-agency structures in "normal" pre-crisis times to routinely question AI-enabled output from human-machine teams.

Policy Brief 4.5 – Identifying Obstacles to Institutional Learning

Researchers:

- Professor Toni Erskine (Australian National University) and
- Professor Jenny L. Davis (Vanderbilt University)

Research Paper: Erskine, Toni and Davis, Jenny L. '"Borgs in the Org" and the Decision to Wage War: The Impact of AI on Institutional Learning and the Exercise of Restraint', presented at 2nd Workshop on 'Anticipating the Future of War: AI, Automated Systems, and Resort-to-Force Decision Making,' 23-24 July 2024, ANU, Canberra, Australia.

Main Argument:

In this paper, we maintain that the state's anticipated reliance on AI-enabled decision-support systems (DSS) in deliberations over whether and when to wage war would be accompanied by a transformation of the state's organisational and decision-making structure, culture, and capacities. If pre-emptive steps are not taken, this transformation could have unintended, detrimental consequences – including in terms of compliance with espoused norms of restraint.

The gradual proliferation and embeddedness of AI-enabled DSS within the state – what we call the 'phenomenon of "Borgs in the org"' – risk leading to four significant changes that, together, would diminish the state's crucial capacities for self-reflection, internal reform and, by extension, "institutional learning". Namely, such reliance on AI-enabled DSS would result in: i) (as least temporarily) disrupted

deliberative structures and chains of command; ii) the occlusion of crucial steps in decision-making processes; iii) deference to computer-generated outputs becoming embedded in the state's practices and procedures, and iv) future plans and trajectories that are overdetermined by past plans and actions. This "institutional atrophy" would, in turn, weaken the state's responsiveness to external social cues and censure, thereby making the state less likely to engage with, internalize, shape, and adhere to evolving international norms of restraint. As a collateral effect, this weakening could contribute to the decay of these norms themselves if such institutional atrophy were to become widespread within the society of states.

Policy Insights and Recommendations:

- **Plan for and repair disrupted deliberative structures.** Introducing algorithmic systems into complex decisions about whether and when to wage war will disrupt existing deliberative structures. As strategic and operational decision-making processes are altered and information flows are re-routed, existing chains of command, along with lines of accountability, will inevitably be circumvented or severed, thereby undermining established processes of review. These changes need to be anticipated so that flows of information and decision-making processes can be remapped, chains of command and lines of accountability re-established, and rigorous review processes maintained.

- **Acknowledge the obstacles posed by obscured decision-making processes**. Even though algorithmic outputs have a documentable quality, which means that organisations can point to these outputs as definitive decision-making factors when reflecting on past action or inaction, these outputs are notoriously difficult to explain, even for those who develop the models. This is especially the case for models that operate through machine-learning (ML) techniques whereby the data processing procedures and parameters of interest are not pre-set, but produced through an on-going interaction between mathematical formulas and dynamic datasets. It is important to acknowledge that specific points on the decision pathway cannot be audited when ML techniques are employed and address the obstacles to institutional review that this creates.

- **Resist embedding automation bias into institutional practices and procedures.** Deference to machine-generated outputs, or 'automation bias', will affect individuals (and teams) at multiple points in the state's decision-making processes over the resort to force. However, an additional, more intractable, problem arises if this human tendency becomes embedded in the state's decision-making practices and procedures. This could happen, for example, by reducing the time allocated for checks on particular predictions or diagnoses that feed into executive decision making, thereby either contributing to a culture in which such checks are viewed as redundant or structurally precluding them altogether.

- **Magnify mechanisms for change.** Formal organisations such as states already have an inherent propensity for stability over change. However, when ML-generated analyses are incorporated into the decision making of formal organisations such as states, resistance to change is magnified. This stems from a basic function of ML processes, whereby the present and future are interpreted by and predicted through training data that derives from the past. The novel interpretations, recognition of new circumstances, and avenues towards change envisioned by human actors must be valued alongside the predictive and diagnostic functions of algorithmic tools.