Transcripts for videos on Ethics of Robotics, Autonomous Systems, and Artificial Intelligence

These videos on the Ethics of Robotics, Autonomous Systems and Artificial Intelligence for Defence (RASAI) were produced by Trusted Autonomous Systems for the Centre for Defence Leadership & Ethics (CDLE) Australian Defence College (ADC).

Find more information on the framework that informs these videos in the Defence Science & Technology Group (DSTG) Technical Report DSTG-TR-3786 'A Method for Ethical AI in Defence'.

Live action videos

Responsibility – Transcript

Narrator: Robotics, Autonomous Systems and Artificial Intelligence (RAS-AI) will provide important asymmetric advantages to Australian Defence. RAS-AI can be used in multiple Defence applications including warfighting, logistics, and humanitarian support.

For all of these applications, humans are legally and morally responsible for decisions or actions that use RAS-AI. That's why Defence has identified responsibility as one of five important facets for the ethical use of RAS-AI. This video is part of a series which unpacks all five facets; these are: responsibility, governance, trust, law, and traceability.

We will delve into aspects of responsibility from the perspective of Defence personnel, including operators and Defence scientists. Using examples from Army use of unmanned ground vehicles and a fictional scenario called Striking Blind, this video will consider human responsibility for the technologies they employ. Through these scenarios, we will come to understand the importance of knowing how a system has been developed, how it behaves, and how to use it. And, we will explore the necessity of providing education and training to Commanders and operators to empower human agency and ensure moral responsibility.

Narrator: The Australian Army recognizes that human-machine teaming offers a potential revolutionary shift in how ground forces plan, train, and fight.

Major Chris Hall has worked on trials that use robotic vehicles for the resupply of combat teams, and unmanned aerial systems for reconnaissance and resupply. He is exploring the applications Army might use RAS-AI for in the future and how human-machine teams might need to be trained to work together.

MAJ Chris Hall: It will be very difficult in the future for a human-only only team to match the lethality and the tempo that can be generated by a human-machine team.

In 2021, Combat Team Charlie experimented with unmanned ground vehicles in a resupply, and logistics roll carrying much of our heaviest kit like ammunition and water. This greatly extended the amount of time we could stay in the field without a resupply. It enabled us to move faster and lighter, and it reduced their signature and the battery charger on the unmanned ground vehicle, also provided power to other systems such as radios and UAS.

The first time we stepped off on a mission within an exercise using an unmanned ground vehicle, within 30 minutes, we had it firmly lodged in a creek line and bogged. It took about 45 minutes to extract the UGV from the creek line. So, that was not a great start to the mission, but it showed that we had not fully integrated that system prior to stepping off on an exercise. And it taught us that the integration of RAS-AI

systems is something that is going to have time and soldiers, and equipment allocated to it before we reach commencement of a challenging mission.

Any tool or weapon can be used unlawfully or unethically, but ADF teams with a strong military ethic will use them correctly, what they currently employ their weapons.

Narrator: Responsibility is a key aspect of legal and ethical use.

To be used effectively and ethically, we must be clear on who is responsible for RAS-AI at all stages of the technology's life cycle from design through to deployment and review.

Responsible use of RAS-AI will require us to draw on expertise from multiple domains including law, to help us govern the use of the new and novel technology.

Lauren Sanders: My name is Lauren Sanders and I am a Doctor of International Criminal Law. I've spent 20 years in the military as both a Signals Officer and a Legal Officer, and my area of expertise is largely Operational Law and the application of International Humanitarian Law to ADF operations. When it comes to responsibility, command responsibility has a legal definition when we're talking about the use of force in armed conflict. So, Commanders have a particular responsibility and the decisions they're making and they have obligations as a part of that as to what information, and what information stand that they need to adhere to, to inform themselves of those particular decisions.

In looking at the use of RAS-AI, the decision making and where it lies along the chain, whether it's the autonomous system making the decision, or whether it's an autonomous system informing a Commander's decision has implications for a Commander and for Defence, more broadly, about how to utilize that RAS-AI system.

Narrator: As Chief Defence Scientist, Professor Tanya Monro is head of Defence Science and Technology Group (DSTG), and Capability Manager for Innovation Science and Technology within the Australian Department of Defence.

Professor Tanya Monro: We need to bring those ethical considerations right to the front of that discovery process. So that we make sure that the future we create through science is something that meet societal expectations and doesn't go beyond what we consider the ethical application of science. The responsibility for AI squarely sits now with everyone both from the point of view of how we use AI but actually more fundamentally from the point of view of our data and how our data is shared and then used in AI.

I put it to you that the responsibility for AI use sits with the individual, as a citizen, and as part of broader society, everyday we create data that is used in AI against us whether it's targeted advertising or any way in which we front the electronic world.

Now in the context of Defence, there's no question that AI gives an advantage and we want that advantage to be asymmetric to give Australia the chance, with our

allies, to prevail in a contested environment. There's a responsibility that sits on Defence to make sure that we have clear standards expectations of the behaviour of both of our civilian and military personnel for the responsible use of AI. That then, of course, had to adhere to International accepted legal norms.

Defence, itself, has an important responsibility in developing the expectations, norms, and ways of using AI. We are keen to get after the capability advantage AI can bring, and that produces a natural enthusiasm but adopting AI too early before we fully understand some of the unexpected consequences, or adverse outcomes of that technology would be rash. So, it is absolutely critical that we work very closely with Defence, the academics, the industry developing these technologies to look at specific scenarios to better understand what we need to do in terms of doctrine and norms of behaviour, to make sure that we both get outcomes from AI but don't have some of the inadvertent, unexpected outcomes that we don't seek.

Also, critically, as we develop AI technology, we need to make sure that ethical considerations are just at the heart of everything we do so that it makes it easier to make sure that AI does what we need it to do, in a way that complies with Legal International norms.

Narrator: Responsibility for critical decisions is spread across multiple decision-makers. RAS-AI can help augment human decision-making and offers advantages in precision and reliability. However, these technologies also have limits.

Any decisions made using AI must be captured using frameworks that indicate ethical and legal responsibility. This will be important in all uses of the technology, but will be especially important when it comes to the use of force.

The complex challenges RAS-AI poses for Defence were explored in a 2021 Perry Group paper called Striking Blind. The story reflected on how decision-makers might be held accountable for acting on flawed recommendations made by Artificial Intelligence.

SQNLDR Sean Hamilton: Sean Hamilton. I'm a pilot with experience in support coordinates, classic coordinates, and a little bit of time flying remote piloted aircraft. The Perry Group is started by Major General Ryan, it initially started looking at science fiction, but it's recently transitioned to looking at what he calls useful fiction. So leveraging off books like Ghost Fleet, written relatively recently to explore future ideas and hypothesize how that might affect us in the business of war fighting.

Our story was Striking Blind and our sponsors from Army Headquarters wanted us to explore the impact of trust in Autonomous Systems. So what trust look like, how you develop trust, and the impacts of having an inappropriate level of trust, either too much trust or too little trust. So this Striking Blind story is about the ADF putting an Artificial Intelligence system called MANDELA into service.

Narrator: The fictional Mandela system is a decision-making tool that provides information about the presence of hostile forces in the battlespace.

SQNLDR Hamilton: Our story is not set in a high-end war. It's set in a low-tech, separatist conflict in the Philippines where we are assisting the Filipino government to re-establish control over areas of territory. MANDELA AI is there with the Ground Force Commander and it's helping identify targets using all-source intelligence so that we can engage threats to friendly forces more effectively. It allows us to get greater speed in decision-making, greater accuracy in decision making. The AI is, in its targeting recommendations, is assessing distinction, proportionality, necessity, humanity, and also the National Rules of Engagement and caveats.

Narrator: The fictional Mandela system is rolled out quickly into an environment that was likely not anticipated by system developers. Mandela relies on information from the environment – including mobile phones and other networks – and enemy forces confuse the system through spoofing and network denial.

MAJ Rebecca Marlow: My name is Bec Marlow, just finished Staff College, and as part of that, I did the Perry Group option and we're looking at AI as part of our Perry Group module. The scenario was an AI targeting system that had been introduced and was being used alongside loyal wingman as part of the Air Force Targeting System, but it was also been utilized by ground-based targeting officers to assist in identifying and verifying targets in the battle space.

The system identified an enemy target, that turned out the target was actually civilians, and that 93 civilians have been killed by this strike.

So, the targeting system had due to the information feed from the enemy had said that this was an enemy, its position and where it was, actually, a civilian position was fleeing refugees from the conflict that was occurring in the scenario.

Narrator: The team who authored Striking Blind were particularly interested in exploring how responsibility for use of RAS-AI might work.

MAJ Dominic Tracey: I'm Dominic Tracey. I'm a Logistics Ordinance Corps Major with a petroleum background, and I was involved in the development of Striking Blind, which was a Perry Group paper, to looking at AI for the future and implementation in the battle space.

This is where the responsibility and the use of AI become more complicated within the military sense. On the surface, it is the pilot who pulled the trigger to engage those targets on the ground, but when you look at the introduction to service, the testing, and validation of that, it becomes a system or a whole of system responsibility and that each key component within the system is responsible for different aspects of that decision, even though their import was well in advance of the decision that was made on the ground. Because AI, when it gets brought into service, has IP issues, learning, how it learns the data that goes into the learning, the articulation of risk for decision-makers to decide to use that system, and the feeds that happen for the pilot.

SQNLDR Hamilton: Does the employment of AI change the Commander's responsibility? We would argue, fundamentally know, in that they are responsible for making sure that the Laws of Armed Conflict and Rules of Engagement are adhered

to, but the extra problem that we've got with putting AI into service is making sure that the Commanders understand when they're expected to trust an AI's recommendations, and when they're expected to question it. Because that could be fairly blurry and we need to clearly document that down, so you would, theoretically, have a higher authority or certification.

Organization that's continually monitoring the AI and giving the Commander very clear policy guidance on when they're expected just take the AI recommendations, and when the AI could be operating out of the bounds that it's certified for and, therefore, the Commander needs to step in and either take the AI out of service or question the decisions a little bit more.

Narrator: The catastrophic error in the Striking Blind story prompts a Senate inquiry.

SQNLDR Hamilton: in the Senate inquiry, initially, it's the viewed as the Commander being responsible because that's our current framework, that it's pointed out by one of the Senators that the Commander had essentially lost free will due to becoming conditioned in trusting the MANDELA AI. We would argue that the accountability should be held at a higher level in the authority that puts their thing into service, the authority that's responsible for continuously accrediting it and making sure it remains battle-worthy, and then the authority for allowing it to be employed in a specific theater, for a specific purpose and period of time.

Narrator: Responsibility is encoded in the law that governs how Australian Defence operates. Damian Copeland is a legal practitioner with expertise in the legal review of weapons and a senior researcher at the University of Queensland. He is a weapons law expert with over 30 years of military service.

Damian Copeland: My name's Damian Copeland. I'm a research fellow at the University of Queensland and part of the Law and the Future of War research team. Our role is to investigate how the law both enables and constrains the use of autonomy in Defence.

The law applies and holds the Commander accountable and others for the lawful use of the AI, which is one of the many tools that may be available to a Commander in any particular situation. If we look at the scenario in Striking Blind, then the question is, did the Commander have sufficient information to enable the decision to be made to authorize the strike in those particular circumstances?

If we look from a legal perspective, the law requires that those who plan or decide upon it an attack to do everything feasible, firstly to distinguish between lawful targets that are enemy combatants, and military objectives and unlawful targets that that is those who the law is designed to protect victims of armed conflict, civilians, and civilian objects. In the context of a scenario, is it sufficient for the Commander to rely exclusively, and on the recommendation of a single tool, an AI tool? Or does the law require the Commander to do more? The answer to that lies in understanding what everything feasible requires in that particular circumstance, so in the scenario, where there other sources of information are available to the Commander?

Could he have called upon the ground forces who he was, ultimately, aiming to

protect to provide more information? Could he have asked the bot aircraft to conduct a another sweep of the area to try and ascertain more information under the cloud level, for example. So these are the sort of questions that are relevant from a legal perspective to determine whether or not relying purely on the AI recommendation is sufficient.

Narrator: Commanders are responsible for how RAS-AI is used. Education and training should be provided so that Commanders and Operators understand the limits of the technology they are using.

The Striking Blind story is a demonstration of what can go wrong when Commanders do not know the limitations of a system.

SQNLDR Hamilton: In this particular instance, the AI misidentified the target for two reasons. One, the insurgents in this particular village had taken away all of the electronic devices from the civilian population. So Wi-Fi, Wi-Fi routers, mobile phones, transmitting devices, which took away one way that the AI could sense that there was civilians present in the area and start mapping population movements.

And then on top of that, the AI didn't notice that all of those devices have been taken away because there was a foreign actor, spoofing, electronic, emissions. And over time, because the AI was sensing that spoofing in every successful engagement at a done leading up to this engagement, each time it successfully engaged a target, in the presence of this spoofing

MAJ Tracey: Within Striking Blind, fundamentally, the issue or the cause of the incident was a lack of understanding and training when it came to the risk of using AI. It came down to not understanding risk of the implementation of AI at the start when it was rushed into service. It came in the training, testing, and evaluation of the AI during its infancy and introduction into service and the training of the pilot in the use of the AI to not understand the risks that are associated with the decisions that the AI has a bias for.

Narrator: When it comes to RAS-AI, Commanders and their operators need training that helps them interpret confidence ratings when the AI categorizes persons and objects.

Corporal Lachlan Jones is an Infantry Signaller dedicated to improving the capabilities and experiences of his fellow infanteers. He is currently in training to become a Systems Engineer and works on a variety of new technologies and data systems that enable autonomous systems to operate.

CPL Lachlan Jones: I'm Corporal Jones, I'm a member of a support company working with 1RAR’s Innovation Space. As with any military equipment or weapon system, the operator is responsible and should be expected to operate within how they've been trained to use the equipment. Soldiers can be better equipped in the moment by well-structured training, comprehensive training, more time on tools, as well as running through scenarios.

Narrator: Responsibility for actions taken by operators, ultimately, sits with the Commanders, whose roles are backed by legal authority.

MAJ Hall: The Commander is always accountable for the actions of the team, but I would draw an analogy between an individual mistake made by a soldier, and a mistake made by a single RAS-AI system. If one soldier in a company commits a negligent discharge, we would hold that Soldier responsible, but if 50 soldiers in a company committed a negligent discharge each, we would say that there is a systemic issue that the Commander should have accounted for in employing their force. Similarly, we don't need Commanders to be able to predict every individual action of every RAS-AI system on the battlefield, but they do need to understand them well enough to be accountable for the results of that system.

The commanders of future human-machine teams in the next conflict that already joined the army and are Lieutenants now.

An officer of the very near future will need to understand the capabilities of an armed UGV or the Air Force's loyal wingman capability as much as they understand Abrams tank or a F/A-18. A good starting point for raising this generation of officers would be the Combat Officers' advanced course and the Logistics Officers' intermediate course. These courses which also integrate for part of their training could implement human-machine teams in simulation and tutes.

Students who have completed that version of the course could then be tracked through to subsequent postings at the Royal Military College Duntroon and Land Warfare Center. There, they would begin teaching human-machine teaming to our trainees and our Junior Officers. Those are students at the college would then flow through having been taught, human-machine teaming in tutes and their field assessments into the Regimental Officers and Logistics officers basic courses where they learn to employ RAS-AI systems in a course-specific environment.

They would be the first generation of officers who could be consistently taught about human-machine teaming throughout their career up until they become Combat Team Commanders. We need to employ a similar train and trainer model in order to raise a generation of soldiers who are comfortable with RAS-AI systems. Starting point would be the implementation of human machine teaming in our Junior Leaders' courses, and a core-specific Subject 2, and Subject 4 courses.

Students who complete those courses would then be tracked through to subsequent postings at Kapooka and our course schools where they would implement human-machine teaming in recruit and course-specific training. Another issue we need to look at in Soldier training is minimizing the training burden, where many systems are very similar. So a soldier might complete a 3-to-5-day course to learn the principles of employing a multi-rotor UAS system, but when they need to transition to a second system, that should be a one-day conversion course.

Technology will continue to evolve too quickly for us to sacrifice our soldiers for multiple weeks at a time every time a new variant is introduced into service.

Narrator: In short, to ensure their own legal and ethical requirements are met,

Commanders and Operators using RAS-AI will need to be confident that the technology will perform as expected, and in line with legal and ethical requirements. Responsibility for RAS-AI is an important issue.

To use these technologies effectively and ethically in Defence, we must be clear on the limitations of different systems. We must also ensure that Commanders and Operators understand the responsibility frameworks that govern their use.

Central to all this is the creation of the ethical frameworks and tools that empower human decision-making and ensure responsibility. Because humans are ultimately responsible for the actions of RAS-AI.

[End]

Governance – Transcript

Narrator: Robotics, autonomous systems, and artificial intelligence (RAS-AI) will be used in many environments in Defence. This might include war-fighting, humanitarian relief, logistics and resupply.

Building legal and ethical governance measures into existing and future RAS-AI capabilities will ensure RAS-AI are fit for purpose and controlled effectively. These governance measures must be guided by our national identity – our values, ethics, and laws – and work with other allied nations' approaches.

Governance is one of the five facets for ethical use of RAS-AI identified by Defence. This video is part of a series on these facets. The five facets are responsibility, governance, trust, law, and traceability.

In this video, we will discuss governance from the perspective of Defence personnel. Governance is concerned with the context technology is deployed in because context determines how RASAI are used and controlled. Using examples from Defence use of autonomous technologies and a fictional scenario called Striking Blind – this video will reflect on how governance mechanisms can act as a safeguard against poor decision-making by both human and machine.

WNGCDR Michael Gan: Hi, my name is Michael Gan from Air Force Plan Jericho. I'm in the Australian War Memorial. I can't think of a better place to talk about the interaction of ethics and technology than where we are right now, which is under the V1 flying bomb from World War II.

You could use an autonomous vehicle, either for attack or Defence, or even to carry ordinance or carry cargo or personnel. So that's within the autonomous systems. You could use artificial intelligence in command and control. One of the applications we're looking at is how commanders make better decisions by taking recommendations of artificial intelligence systems that can gather enormous amounts of data and bring out really important things for the commander.

You could also use artificial intelligence for force protection such as for medical purposes. One of the real effective uses of artificial intelligence is to help doctors identify and diagnose issues by taking a large amount of information from their patients, such as scans on identifying things like lung problems, cancers, tumours, and similar things like that.

Narrator: RAS-AI promise game-changing advantages for Defence. But they also come with ethical and legal risk.

WNGCDR Gan: When we choose to use these applications in the military context, in conflict, in the gray zone, or close to conflict, we do run into a lot of ethical issues. There are a lot of things that we have to consider that they even considered back in World War II with the new weapons as technology was brought in, in each of those applications.

For example, in the command and control situation. The artificial intelligence, which is supporting the commander in his decision support, mustn't be biased. It must know what information the commander wants and doesn't want. But most importantly, the artificial intelligence should have value alignment. So it should have the same values as the commander, the government, or the force which it's working for. So that's a key and a key thing of importance.

There are a lot of critical areas we have to consider. The application of ethics with the use of robotic and autonomous systems in military and in conflict zones.

Narrator: We must have methods for governing and controlling RAS-AI to ensure Defence can develop, acquire, and deploy these technologies in an effective, ethical, and lawful manner. Governance can be provided in multiple forms including law and policy, instructional guidelines, safety manuals, and culture. The Australian Ethics Doctrine is one example of this.

MAJ Dominic Tracey: I'm Dominic Tracey, I'm a Logistics Ordinance Core Major, specializing in fuel. And then I was involved in the Perry Group Paperwork Striking Blind, which explored AI and its introduction to the battlespace

MAJ Tracey: In my mind, for use and development of RASAI into the future, the Australian Ethics Doctrine provides a framework in which we can start testing and evaluating AI decisions and more importantly start seeing how we can start shaping data going into the AI to then come out with appropriate decisions for the ADF into the future. And it just sets the parameters for the AI for what is acceptable and not acceptable into the future.

Narrator: Governance mechanisms impose controls on RASAI throughout the life cycle of the technology.

Australian Chief Defence Scientist Professor Tanya Monro has said that AI technologies offer many benefits such as saving lives by removing humans from high threat environments and improving Australian advantage by providing more in-depth and faster situational awareness.

Governance frameworks from RAS-AI will help secure those advantages while ensuring better, ethical decision-making. It could also improve how we work with our allies.

Professor Tanya Monro: In the field, commanders and individual soldiers and ADF members always have been and always will be responsible for their own decisions. But of course, this gets richer and more complex when we bring AI into the picture. We need to make sure that we really understand how those decisions, how those instructions made by individuals and commanders in the field lead to outcomes through AI. The key here is making sure that we have a good working understanding, fully implemented of the international legal norms, and that we focus on interoperability with our allies. Because taking different approaches across elements that are meant to be interchangeable and interoperable in an allied warfighting environment could have disastrous consequences, which really pushes us into thinking about interoperability and interchangeability in the field in quite a

different way than we have before.

In the past, it's been much more about do our war-fighting platforms work well together? Can they communicate? Can they operate side by side to achieve a shared outcome? Now it's a much deeper question around data, and what's done with that data, and how that data is shared. And then, how that data is used to guide in the moment decision making. I think it pushes us to a deeper embedded interaction with our allies, which I think will give us capability advantage back in 2019. Defence led a workshop that brought together some of the most expert individuals from around our nation in areas as diverse as specific areas of AI and machine learning, law, ethics, and Defence to really look at what we needed to form the foundations of a framework for ethical use of AI in Defence. The outcomes of that work and that report, I think, set the foundation, the context for us developing that capability.

It's not about having a high-level abstract statement of intent. It's much more about building from the ground up and understanding of both the principles and the constraints that need to be put around AI and its use in the field. There's no question that the legal implications of the use of AI is going to create some very, very interesting case studies into the future. I'm going to start this part of the conversation by looking at Australia. Australia has really clearly articulated obligations as a signatory to international laws on military activities in weaponry. Our obligations under those are clear and well understood. But of course, were developed in a pre AI world. As we move forward and understand more fully the potential uses of AI through some of the work we do in applying it to some of our priority challenges, we won't escape the need to apply and use and comply with that legal framework. But what we will uncover is where we need to work with allies to refine, develop and clarify that framework as it applies to our actions.

One thing that I think goes as a central tenet of our work with allies is that we work in a way that complies with international regulations, and we don't try and hide beneath the lack of clarity that laws that have yet to catch up with technology might have. Now the increasing role of non-state actors and terrorist organizations or criminal networks or militias does pose an additional challenge because they're not constrained to the same norms. Nevertheless, Australia and our allies will comply to norms. But we can focus on creating asymmetric advantage while still complying with those norms in order to deter the kinds of behaviours that will destabilize international order and stability. Because democracies simply cannot afford to breach that trust with their citizens. So this is a deep value we share with our allies.

Narrator: Human-machine collaboration should be optimised to safeguard against poor decision-making. This collaboration can be supported by governance frameworks that ensure creators understand the context in which RASAI is used and how the technology is controlled.

In the 2021 Perry Group Paper Striking Blind, Australian War College students used science fiction to consider the challenges of military AI and autonomous systems set in 2040. Their paper depicts the use of a fictional artificial intelligence system called Mandela, which is deployed too quickly and without transparency into an environment where it is not fit for purpose.

MAJ Rebecca Marlow: My name is MAJ Bec Marlow. Just finished Staff College and as part of that, I did the Perry Group option. We're looking at AI as part of our Perry group module.

We wrote our papers part of Perry Group module looking at some of the issues that could be faced by and challenges that could be faced for future AI procurement within Defence. It was one of the questions that was posed by the CDF.

The scenario was an AI targeting system that had been introduced and was being used alongside Loyal Wingman as part of the Air Force targeting system. But it has also been utilized by ground-based targeting officers to assist in identifying and verifying targets in the battle space. In our paper, it was an American based system that we had been asked to introduce sooner rather than later due to competing system being utilized by Chinese bosses. So it became important that we had a near peer level of ability to target. We were still introducing it into service when it actually became critical that we employ it. So, not all the testing and verification had been completed on the system by the time it started to use as part of that paper.

Narrator: AI systems should be deployed only after demonstrating effectiveness through experimentation, simulation, and live trials. Robust testing is necessary so that we can assess AI decision making in relevant scenarios.

MAJ Marlow: So that was what we were looking at, what happens when you don't go through the complete process in order to introduce a system into service or start using a system before the full introduction into service piece has been completed.

It uses and drew data from the environment. So, from mobile phones or networks, other networks in order to inform what the targeting picture was and whether or not the target was a viable target or not. In this case, we were actually one being spoofed but also there was network denial which affected the ability of the targeting system to actually correctly identify and then prosecute a target.

Narrator: Even though the fictional AI has not been fully reviewed, MANDELA is deployed in a complicated noisy environment to improve decision-making and support operators. However, the data used by the system is flawed and things start to go wrong.

MAJ Dominic Tracey: Within Striking Blind, our main protagonist Charles engages ground targets based on a feed from the AI saying there is a legitimate target. What leads up to that is a ground commander calling for fire support. The AI system identifies a group of targets that are legitimate targets and they’re engaged. What it doesn’t do is translate that into a secondary target very well. So, the same pattern of behaviours identified through a secondary group, which turns out to be a civilian group. And then the same pilot is given the same information feed, makes the same decision based on that information feed. However, it strikes illegitimate target and causes unacceptable collateral damage. And then the rest of the story talks the fallout and explores the issues because the fallout of that decision,

Narrator: We need to make sure that RAS-AI enable better, more ethical, and lawful

decisions and actions. To achieve this, operators need to be awareness of a systems' actions, decision, behaviours, and intention. And training is needed to ensure that human-machine collaboration is optimised to prevent automation bias and over reliance on the machine.

MAJ Tracey: In the Striking Blind scenario, Charles was conditioned to have confirmation bias to trust that the AI's recommendations were going to be correct and that there were valid. Unfortunately, during the training that he received, he was never required to question the AI or to reinforce what scenarios he needed to question the AI more than other scenarios, which then conditions people to trust or overly trust the AI's recommendations rather than to critically assess them. One of the problems when it comes to AI is that AI increases the speed of decision making or has the potential to do so, but the human still needs to critically evaluate that. If you just rely on a human critically evaluating every single decision, then you nullify the benefits of AI in that speed of decision-making cycle.

Narrator: Governance covers large range of methods for supporting and constraining the use of RAS-AI in Australian Defence. To ensure a safe and effective use of RAS-AI, these methods should work together and provide redundancy measures.

The novelty of these emerging technologies means that we may need new ways of governing human decisions' in complex and complicated environments. This will include legal control measures that could be integrated into the system during design phases.

Damian Copeland is a Senior Research Fellow in the University of Queensland Law and Future of War group. Damian's research focuses on the application of export control, arms trade and sanctions regimes relevant to the export and brokering of trusted autonomous military systems and associated technology.

Damian Copeland: Any technology is going to have limitations. So, any operator who is responsible for the use of an AI system must understand its capabilities, but also the limitations and therefore be able to use that knowledge to use the AI effectively. This will be gained from a thorough test evaluation and validation process which will endeavor to identify circumstances where the AI is not able to perform to the required standard where the AI susceptible to vulnerabilities, particularly from enemy actions or enemy's attempts to confuse the system.

So, before commanders actually given an AI capability to use in armed conflict, there needs to be a thorough understanding of what those limitations are before a command is actually given an AI capability to use in armed conflict. There needs to be a thorough understanding of what those limitations are. And then, those that are responsible for its use both commander and operator levels should be thoroughly trained and practiced in the use of AI in the environment where it's anticipated to be used. This can be expressed in the tactics, techniques and procedures that are developed for the use of the particular system. This can also be the product of recommendations from an article 36 review, which could recommend that in certain circumstances, the use of AI should be limited or prohibited to ensure that it's capable of being used lawfully.

Narrator: Article 36 reviews are part of Australia's legal obligations, such as responsibilities imposed by the Geneva Conventions. These laws set limitations on the ways war may be conducted.

Article 36 reviews evaluate technology throughout design and development; providing an opportunity to review and improve compliance and secure more humanitarian outcomes.

Major Chris Hall has worked on trials that use robotic vehicles for the resupply of combat teams and unmanned aerial systems for reconnaissance and resupply. He is working on ways to prepare future army leaders to work alongside machines and human machine teaming.

MAJ Chris Hall: As RAS-AI systems develop further in complexity and with lethality, it will raise issues that are specific to raise RAS-AI systems. There's the technical aspects such as whether a system can distinguish between a farmer's vehicle or a technical vehicle with a weapon mounted on the back.

More broadly, there are larger ethical considerations. The implications of RAS-AI systems for commanders to be able to ethically employ them are not necessarily obvious if you haven't worked in this space. I've had to work pretty hard to understand some of them. What would be really useful is case studies of ethical and unethical employment of these systems to really demonstrate in a tangible way to commanders how it can go right or wrong.

Any tool or weapon including RAS-AI systems can be used unlawfully or unethically. But any of the teams that have a strong military ethic will use them correctly like they currently employ their weapons. As the lethality and complexity of RAS-AI systems increases, more issues may come specifically from employment of RAS-AI.

Narrator: RAS-AI can help enhance the way Defence works. However, to be used effectively, new technology must be accompanied by comprehensive training packages to help Commanders and operators know the frameworks that govern their use as well as how to use the systems at hand.

CPL Lachlan Jones: I'm Corporal Jones, I'm a member of a support company working with 1RAR’s Innovation Space.

With RAS-AI I can automatically process data. It may place a low significance on potentially important information that could change the outcome of a mission.

Operators need a comprehensive training package to enable them to operate RAS-AI in complex and stressful situations. Operators need to be qualified and experts on the equipment. They need to know their SOPs, TTPs, their rules of engagement and the laws of armed conflict to effectively operate the machinery.

Narrator: We need new and emerging technology to be aligned with those values in our legal obligations. RAS-AI will be ethical where the humans designing, developing and using the system are guided by those principles.

MAJ Tracey: If Australia was involved in the initial phase of design or development of that AI then some of our morals, values and ideals could be implemented, which is one of the things that Striking Blind is trying to do. It is to prompt the conversation that Australia needs to be involved in AI development so that the product that is presented later on in 2044 for us in the near future is appropriate for us to use rather than relying on another countries or another organization's morals and values opposed to Australia's morals and values. Within Striking Blind, fundamentally, the issue or the cause of the incident was a lack of understanding and training when it came to risk of using AI. It came down to not understanding risk of the implementation of AI at the start when it was rushed into service. It came in the training and testing and evaluation of the AI during its infancy and introduction to service, and the training of the pilot in the use of the AI to not understand the risks that are associated with the decisions that the AI has a bias for.

Narrator: Returning to the fictional Striking Blind scenario, the RAS-AI system might have been deployed used to better effect if improved governance frameworks were in place.

SQNLDR Sean Hamilton: I'm Sean Hamilton. I'm a pilot with experience on Super Hornets classic coordinates, and a little bit of time flying remote piloted aircraft.

What sort of governance structures would have helped manage the employment of Mandela? A couple of different issues to think about. The first one is not having a clear understanding on how exactly Mandela is making decisions. That is a technical and legal problem with how we are going to buy artificial intelligence systems from the United States and have a clear understanding of how exactly they're developed, how they're taught, how they're learning on the fly, and how can we analyze their decisions and pick up any errors that happen in the course of operations. So that's going to require very close work with industry.

We may need to be inside the tent when they're developing these systems and we need to have really clear arrangements for the flow of information. We need to make sure that when we purchase the artificial intelligence system, it's not just buying at the shop and taking it home. It's a full life cycle approach where the industry that made the system is fully embedded in the operation of the system and they're helping the ADF continue to monitor its functionality, conducts continuous tests and making sure that it's still behaving in the way that we expect it to behave so we've got full disclosure on how exactly it's making decisions.

The other part of the problem is the accountability framework. So in our story, there's no accountability framework beyond what we currently have at the moment where the buck stops with the commander. And the ground force commander that's getting these recommendations is fully responsible for making sure that we adhere to ROE and the laws of armed conflict. That was inappropriate in this particular instance because he was put in a position where he was conditioned to trust the AI and essentially, it wouldn't have been reasonable for him to overrule that decision.

So we need to put more thought into the accountability structure. Who's accountable in this chain? One could argue that we would need a body or a person, a

commander up much higher to authorize the system into service and then make sure that it's specifically authorized for that theater and give very clear guidance to the commanders on when they're expected to just take recommendations, and they'll be backed up legally in taking those recommendations, and when they are expected to question those. What sort of left and right bounds of assumptions should the commander be operating?

Narrator: Different elements in our existing structures may have to be adapted in order to support governance of RAS-AI. For instance, intellectual property and international traffic in arms regulations frameworks.

MAJ Marlow: Intellectual property that is the software system in the programming inside the Black Box belongs to the company that created it. So, having access to that, if it needs to be modified by us to suit different operating scenarios because we discover through validation and verification process that we are using it in a slightly different way or with slightly different parameters to how the U.S. was employing it, we need that ability to be able to change it. ITAR is the International Trade and Arms Armaments Regulations and that is imposed on sale and use of arms and technologies by foreign forces.

So there are restrictions on what we can know within a system, our ability to understand what was inside the system. So, whether that is purchasing the IP from the company or creating a legal framework with the United States that better enables us to work within the ITAR framework to then get access to what's inside the system to better understand that. Because if they were the ones who asked us to use it because of the near-peer advantage with the Chinese, then there needs to be some framework allowed to enable us to actually get inside and have a look, and then adapt it to what is available for the Australian system to use.

Narrator: Governance methods provide protection from poor decision-making by humans and machines.

Control measures will be critical to ensuring that RAS-AI capabilities can perform their functions, both lawfully and ethically.

Commanders and operators will require training in both capabilities and limitations of the technology as well as the measures we create to control them.

[End]

Trust – Transcript

Narrator: Robotics, autonomous systems, and artificial intelligence (or RAS-AI) will introduce game-changing new capabilities for Australian Defence.

Defence has identified trust as one of five facets for ethical AI because trust will play a significant role in how these technologies will be used. This video is part of a series on these facets, which are responsibility, governance, trust, law, and traceability.

In this video, we will explore trust in RAS-AI from the perspective of Defence personnel using case studies from Army use of unmanned ground vehicles and a fictional scenario called Striking Blind.

Trust is a belief in reliability or ability of someone or something. It is vital for the adoption of any technology. For RAS-AI, trust will exist where the actions of systems are coherent and justified.

Trust requires hands-on experience in training and exercises to make sure our people and machines are trained and prepared for real-world scenarios. In the Australian Army, Major Chris Hall has worked on trials that use robotic vehicles for the resupply of combat teams, and Unmanned Aerial Systems for the reconnaissance and supply.

He is examining the best ways to prepare future Army leaders to work alongside machines and looking at human-machine teaming more broadly.

MAJ Chris Hall: I've been convinced that these systems are the way forward in achieving increased lethality and protecting our own people. It will be very difficult in the future for a human-only team to match the lethality and the tempo that can be generated by a human-machine team.

Human-machine teaming, and the use of RASAI systems are not currently taught within the all-core officer, or all-core soldier training continuums.

One of the best ways we could set up our teams for success in the future, in making ethical and lawful decisions with these tools, is to give them years of experience prior to deployment.

Narrator: In the Australian Army, Corporal Lachlan Jones is working on developing robotic and autonomous systems. As part of this work, he developed an app for a system that helps soldiers send reports.

CPL Jones: I'm Corporal Jones. I'm a member of a support company working in the 1RAR innovation space. So for the last 12 months, I have been working on photogrammetry and point cloud data that can be generated and disseminated on the Tactical edge of the battlespace to enhance section commanders level of awareness. Through the production of 3D models and the information generated throughout the can be disseminated on the battlespace to enable section commander's greater situational awareness.

Narrator: Defence personnel are now working on how these technologies can be

safely integrated into human-machine teams.

However, before we use robotics autonomous systems and artificial, we need to establish a baseline of trust in these technologies so that commanders and operators can be confident that a system will perform as expected and in line with governance requirements. This will help Defence personnel ensure their own legal and ethical obligations are met.

Trust or knowledge of the reliability of the technology is therefore essential.

At the Australian company of Athena AI, trust is a central component to the technology they develop.

Stephen Bornstein: I'm Stephen Bornstein. I'm the CEO of Cyborg Dynamics Engineering and I'm the managing director of Athena Artificial Intelligence, which is a spin-out AI company. Initially created through the justified autonomous UAS project within the TAS DCRC

My personal background, I spent 10 years doing research and development in the robotics AI and Defence engineering space. Working for BAE Systems, Electron launch vehicle, Rocket lab, Airbus, helicopters for MRH 90 entry into Special Forces. And now with Cyborg Dynamics. Also, an Army Reserves infantry officer as a reservist, I spent five to six years doing that as well.

We do artificial intelligence to support our lawful targeting on the battlefield with Rules of Engagement for human-in-the-loop operations. And that works with ground Robotics, air Robotics all the way up to strategic level assets.

I think it's the most important thing for us, it's very easy to produce a Artificial Intelligence detector. And then you can combine that detector with a camera and you hall have some sort of result. But how do you know whether you can trust that footage to actually only send the detection off the battlefield where you need assurance. Assurance is associated with trust and we need to consider how we actually do that.

Narrator: The 2021 Perry Group Paper Striking Blind considers the opportunities and limitations of military Artificial Intelligence and autonomous systems.

In this story, human operators become conditioned to trust the decision-making system with opaque algorithms, which was implemented after a limited review and limited training for Commanders and operators who lacked sufficient knowledge around the reliability of the system. Striking Blind illustrates why trust requires full knowledge of a system.

SQNLDR Sean Hamilton: I'm Sean Hamilton. I'm a pilot with experience on Super Hornets classic coordinates, and a little bit of time flying remote piloted aircraft. Our story was Striking Blind and our sponsors from Army Headquarters wanted us to explore the impact of trust in autonomous systems. So what trust looks like, how you develop trust, and the impacts of having an inappropriate level of trust. Either too much trust or too little trust.

This Striking Blind story is about the ADF putting Artificial Intelligence system called Mandela into service. Mandela - think of a room size Artificial Intelligence system, theoretically, it's a decision AI. So it helps commanders make targeting decisions in our store. It was developed by the United States to fight a high-end war against China, because the high-end conflict decisions were so rapid and complicated that humans were not able to make them appropriately.

So they field an AI. Australia bought that AI under pressure from the United States with the US highly recommending that Australia, purchase the AI and put into service to facilitate integration and interoperability in coalitions. In addition to that, Australia was doing modelling looking at their likelihood of success in a war against China. And realized that they will be at a significant disadvantage if they are purchased the AI. So there was a lot of pressure to go ahead and field it.

Fast forward, our story is not set in a high-end war, it's set in a low-tech separatist conflict in the Philippines, where we are assisting the Filipino government re-establish control over areas of territory. Mandela AI is with the Ground Force Commander and it's helping identify targets using all-source intelligence so that we can engage threats to friendly forces more effectively.

It allows us to get greater speed and decision-making, greater accuracy in decision making. Their AI is in its targeting recommendations is assessing distinction, proportionality, necessity, humanity, and also the national rules of engagements and caveats.

Narrator: Striking Blind is a story centred around a crew of a Super Hornet, which are airborne over the Philippines, escorting a Filipino special forces patrol located on the ground. The Mandela AI is working alongside the crew, monitoring for hostile people.

SQNLDR Hamilton: The Mandela AI just recommends targeting suggestions to the commander. And then the commander on the ground is target engagement authority is ordering the strike.

So he goes ahead and order the strike, 32 enemy fighters are destroyed on the ground and the Special Forces Patrol is protected. That engagement is a success the fighters of then re-rolled to a second engagement of Bridge seven kilometers to the South. Where our sensors have picked up seven technicals on a bridge which Mandela then deems hostile because they're an imminent threat to their special forces Patrol.

The AI thinks that they are about to step off and commence an engagement. The commander gets that decision. They attempt to get eyes on those vehicles, but due to low cloud they can't. So ultimately Commander weighs the risks and accepts their Mandela engagement recommendation based on it, being an imminent threat to the friendlies. Weapons to drop through the cloud and at the time successfully destroyed the technicals.

The story fast-forwards four years and it has now come out that engagement was a

failure and it was one technical on a bridge. That was a checkpoint stopping civilians fleeing a village and the strike killed 92 people and injured another 100 or so.

Narrator: The story of Striking Blind was created to illustrate the risk of trusting a system which is not fulling understood.

MAJ Dominic Tracey: I'm Dominic Tracey, I'm a Logistics Ordinance Core Major, specializing in fuel. And then I was involved in the Perry Group Paperwork Striking Blind, which explored AI and its introduction to the battlespace. So, the Perry Group was an initiative under Major General Ryan within this ACSC to look at future concepts and explore that through science fiction. So it was a way to engage more readers and introduce problems that may not interest people on a day-to-day basis.

So within the AI system of Mandela that we had for Striking Blind, it had the standard kill chain process over the top of it, with a human in the loop at the final end to decide to engage a target. What it lacked was some of the fidelity and its introduction to service where it was rushed into service to meet a capability gap, that was generated through near-peer competition at the time. And that because a Robotics introduction to service didn't happen.

Some of the test cases that should have been explored further to tease out risk were not done. Which then leaves decision-makers with a gap for how it should be mitigated against all those risks should be mitigated against.

Narrator: Human in the loop refers to people working with the RAS-AI system to change the outcome of an event or process.

In the story, the human was insufficiently trained in the use of the system and does not understand its limitations.

MAJ Rebecca Marlow: My name is Bec Marlow. Just finished Staff College and as part of that, I did the Perry Group option. We're looking at AI as part of our Perry group module.

So not all the testing and verification had been completed on the system by the time it started to be used as part of a paper. So that was what we were looking at was what happens when you don't go through the complete process in order to introduce a system into service. Or start using a system before the full introduction into service.

So we did have a human in the loop. So the targeting officer on the ground, was the person who in the end, made the decision on whether or not the AI had made the correct decision. And in this instance, with a lack of other available data from other feeds because we were in a denied environment. They trusted the system because he had previously been correct. And they trusted that this time again, even though we didn't have all the feeds that it had the feeds and it was correct and therefore allow the system to identify and then fire on the target.

Narrator: The catastrophic mistake made in the Striking Blind story prompts a review into how the system was used and relied upon by the Australian Defence Force.

SQNLDR Hamilton: There is now an inquiry into how exactly the ADF have got into this position where we're just taking recommendation, straight out of been AI. Largely without questioning them and too much detail and this is resulting in significant collateral damage. That investigation in the Senate turns out a range of issues. So it covers off on why Australia bought the AI in the first place. Like, they were pressured to put it in the service, due to interoperability reasons and maintaining a combat effective Force.

They would have like to have had a greater understanding on how exactly the AI was processing its decisions. But they were prevented from gaining a really high level of understanding due to International traffic volumes regulation, preventing information flowing to Australia, intellectual property rights. But then also technical limitations in figuring out how exactly an AI that's doing machine learning is actually making its decisions and then adapting its decisions.

The last problem that they came to is that they put Mandela into a semi-automatic mode, were was recommending, engagement decisions to a commander and the commander was making the decisions. And so the human was in the loop there to try to mitigate the risk of having an AI engaging targets without the human in the loop.

But what they had found is because Mandela was so successful. So often and had recommended successful engagement so many times the human in the loop, the commander and now become conditioned to trust it. When presented with that scenario where there's, he's told that there's seven technicals on a bridge in their imminent threat to the friendly. Mandela calls them hostile Mandela has called a lot of things hostile in the past and it's proved successful. Why would he not take that recommendation? It would be unethical for him to not act on that information, because of the imminent threat to friendly forces in the balance of probability that Mandela's got it, right.

So having the human in the loop was completely ineffective. It wasn't generating the risk management that the ADF wanted because the Commander's had become conditioned to trust it. If they went to the loop, it would have had the exact same outcome with Mandela autonomously recommended engagement.

Narrator: Striking Blind is a useful illustration of why trust with verification is vital when it comes to acquiring and deploying RAS-AI.

To be trusted RAS-AI needs to be safe and secure. It must operate reliably in accordance with its intended purpose and must provide a useful level of transparency and explainability.

We must make sure the RAS-AI is trusted ethical and transparent before they bought into service.

As Chief Defence Scientist, Professor Tanya Monro is head of Defence Science and Technology Group (DSTG), and Capability Manager for Innovation Science and Technology within the Australian Department of Defence.

Prof Tanya Monro: We have got a number of layers we need to worry about, you know, human in the loop, human on the loop, and humans out of the loop. They're all different levels of automation. The way we get trust is through transparency and verification. And one of the key focus areas for this work has to be delivering methodologies that are agreed and accepted for delivering that transparency and verification.

What information do we as humans need to see from our AI algorithms in order to get that confidence that the data is producing meaningful outcome? It's interesting because when we reflect on these two core foundational principles of trust, it actually makes you reflect on what you need to do to drive the whole AI Enterprise for Defence. Because we know that to deliver some of the asymmetric advantages and to get humans out of perilous situations. We need to work very deeply with industry and with Academia. But we can no longer set some requirements and sit back and wait for industry to deliver an outcome.

What we really need is to work in deep partnership to make sure the principles of transparent verifiable AI a threaded right through both from concept development. When we generate knowledge through our academic partner work. Through to the delivery of solutions and products that are commanders and ADF will use in the field. Just as we need the AI itself to be transparent and verifiable. We need our solutions to deliver that as well, and that must be ingrained in the way we work as an ecosystem.

Narrator: Working with industry during the development of RAS-AI could help Defence understand the values and assumptions that go into the creation of a system. In the Striking Blind scenario, blind trust and a lack of understanding of the system, resulted in a catastrophic error.

MAJ Marlow: In our scenario, the system at always being correct. We did make it so that they had trust the system. And that was a part of building that as they made the decision that they made. Because previously, the system had always given the right information and given the right targeting information, so that we were able to trust her. When they didn't have the full information, they trusted that the system did have the full information, which is why they trusted that when they did despite the outcome.

MAJ Tracey: So at the other end of the spectrum, when it comes to our AI, the uses of AI, they need to test that AI to failure. So in the Striking Blind scenario Charles was conditioned to have confirmation bias to trust that the AI is recommendations were going to be correct and that they were valid. Unfortunately during the training that he received. He was never required to question the AI or to reinforce that what scenarios he needed to question the AI more than other scenarios.

Which then conditions people to trust or overly trust the RAS-AI recommendations rather than to critically assess them. One of the problems when it comes to AI is AI increases the speed of decision making or has the potential to do so, but the human still needs to critically evaluate that. If you just rely on a human critically evaluating every single decision, then you nullify the benefits of AI in that speed of decision-making cycle. Within Striking Blind that the trust Dynamics between the pilot and the

AI to push that point even further.

The pilot needs to be conscious of every time that they are just accepting what the RAS-AI giving them without critically evaluating. And within that not only the pilot, but the entire kill chain involved, also need to be able to look at the data that AI has recommended their decision based on within Striking Blind without looking at the learning loop of the AI.

If it has engaged a target, as then being confirmed to be a legitimate target based on a pattern of behavior. If that AI learns within that short period of time, but the rest of the system or the humans involved in that system have not learned in that time or don't understand how that AI has reinforced that decision making, then they can't test and adjust future assessments. Which is what happened for that second strike.

Narrator: Trust is a relationship comprised of competency and integrity. Competence requires skills, reliability, and experience. Integrity requires good character and professional competence. When RAS-AI is used in Defence, operators will hold different levels of trust in a system depending on how the technology fulfills those components of competency and integrity.

To evaluate the competency of a system, we need to understand its limitations. Legal review can assist with this.

Damian Copeland is a Senior Research Fellow in the University of Queensland Law and Future of War group. Damian’s research focuses on the application of export control, arms trade and sanctions regimes relevant to the export and brokering of trusted autonomous military systems and associated technology.

Damian Copeland: Any technology is going to have limitations. So, any operator, who is responsible for the use of an AI system must understand its capabilities, but also the limitations and therefore be able to use that knowledge to use the AI, effectively. Will be gained from a thorough test evaluation and validation process which will endeavor to identify circumstances where the AI is not able to perform to the required standard. Where the AI is susceptible to vulnerabilities particularly from enemy actions or enemies attempts to confuse the system.

Before a command is actually given an AI capability to use in armed conflict. Then there needs to be a thorough understanding of what those limitations are. Those that are responsible for it's use. Both a command and operator levels, should be thoroughly trained and practiced. In the use of the AI in the environment where it's anticipated to be used.

Narrator: It is critical that personnel has trust and confidence and the systems they are fielding. Commanders in particular, should adopt a trust with verification approach to ensure that RSA-AI are used ethically and lawfully.

CPL Jones: I think trusting something that can potentially make decisions without any human input can be hard but now three working with autonomous programs that collate and process data are now trust those to process data without any input. Whereas originally, I would constantly track them and monitor them. Soldiers, and

people in general, tend to distrust things that they don't know about. So time on tools are plenty of instruction as well as running through constant continuation training, can help build levels of trust.

Narrator: It is critical that Personnel have trust and confidence in the systems they are Fielding. Commanders in particular should adopt a trust with verification approach to ensure that RSA-AI are use ethically and lawfully.

MAJ Hall: Commanders are not going to employ tools in a high-risk mission, which they do not trust. With the technology available to us now where I would trust the RAS-AI system is in logistics and resupply roles or reconnaissance. Especially under the control or supervision of a human being. I would also trust that the vision from a RAS-AI system can be used by humans to make good targeting decisions. In contrast, we know that with the tools available to us now Soldiers are required when we need to directly attack the enemy.

The best support that we can provide to soldiers and commanders in using RAS-AI systems is experience and familiarity before they get to a challenging mission. So we need to get these tools into the hands of soldiers and commanders in barracks and on exercise and that will make them most likely to make ethical decisions in the same way that currently do with their weapons. Like any tool the team will begin to trust RAS-AI systems after they have seen consistent behavior and worked with that system.

Narrator: Trust is essential for people working in Defence because they are trusted to do things that are otherwise restricted in society.

Similarly, RAS-AI used in Defence needs to be trusted to work in high-stakes contexts. To be trusted, RAS-AI systems need to be safe and secure within our Nation’s sovereign supply chain. They must also operate reliably in accordance with their intended purpose.

Operators will hold multiple levels of trust in the systems they are using depending on which aspect of trust is under scrutiny. Trust may change over context and time and must be proportionate to the risks.

Trustworthy RAS-AI must be lawful ethical and robust. Commanders and operators must be provided with pragmatic training that establishes how and when RSA-AI can be trusted and when they should verify it. Trust can be established through rigorous tests and evaluation, and then extended through hands-on uses of the technology in exercise and operations.

[END]

Law – Transcript

Narrator: Robotics Autonomous Systems and Artificial Intelligence or RAS-AI will deliver game-changing capabilities for Defence.

RAS-AI are lawful when they are capable of performing their function, in compliance with the operator's legal obligations. Law is one of five facets for ethical AI identified by Defence. This video is part of a series on these facets, which are: responsibility, governance, trust, law, and traceability.

In this video we will review the facet of law from the perspective of Defence personnel using a fictional scenario called Striking Blind to illustrate the importance of our legal frameworks, which review and evaluate new means and methods of warfare.

In Australian Defence, we review the legality of all new means and methods of warfare, through what is known as an Article 36 review. Legal reviews are informed by Australia's commitment to International Humanitarian Law, which requires measures that reduce adverse humanitarian effects that result from warfare.

Lauren Sanders is a legal practitioner with over twenty years of military experience, with expertise in International Human Law, including advising on the accreditation and use of new and novel weapons technology.

Dr Lauren Sanders: My name is Lauren Sanders. I am a Doctor of International Criminal Law. I've spent twenty years in the military as both a signals officer and a legal officer. My area of expertise is largely operational law and the application of International Humanitarian Law to ADF operations.

I'm also the managing director of a small legal firm called International Weapons Review or IWR. We focus on providing industry advice as to how they can operationalize their capability, focusing on legal compliance.

A great example from recent operations would be the use of ISR to augment the targeting capability, or the visual range of an Apache helicopter in operations in Iraq in 2017. Then MUM-T, but now HUM-T, a Human-Machine Teaming, which was using the capability of the reaper, to effectively act as a forward location to visualize what the pilots were looking to target.

Interestingly, because that system was one that was being used or tested in operations, the level of trust by command wasn't there yet.

What was actually happening instead, is that those systems were being used to assist in target verification, only after the targets had been verified through the traditional deliberate targeting process.

Hopefully, a mature version of that system will be using that HUM-T process. To actually speed up and extend the range of those capabilities, without having to come back to a Command Decision Headquarters. To actually go through that deliberate targeting process, but use it more as a dynamic targeting system.

The legal and ethical issues that were foremost in my experience during the introduction of these UAS, was two specific areas.

The first was, whether or not the capability complied with our legal obligations to actually use in the first place. And part of the process of introduction to service of new capabilities, particularly where they're going to have some sort of kinetic effect or connection to a kinetic effect. Requires that those capabilities undertake what we call an Article 36 review, which is an article of additional protocol one to the Geneva conventions, which requires that methods, means, or weapon systems, are actually checked to comply with the International Law Obligations.

Prior to the introduction of the UAS, they were checked to see if all of the system's capabilities and additional bits and pieces that were associated with them, actually complied with the International Law requirements.

Thinking about some of the additional component tree of the UAS that we wouldn't necessarily have thought about before using it.

That was a process that needed to be tested, assured, and then systems adjusted to make sure that, that wasn't a problem. And therefore we could use them in compliance with our legal requirements.

The second issue that came up when we were talking about the use of UAS and its introduction, was really that command trust perspective. And whether or not a commander who was making a decision to rely on an information feed coming from the UAS, would satisfy their requirements. And from a legal perspective, satisfy their legal standards to make a decision about targeting.

Commanders have obligations as decision-makers in that targeting cycle. They're ultimately the individuals who are responsible for the decision to release the weapon system. So, when they were working through the process of integrating those ISR feeds from various different locations into the targeting cycle, there was a period of adjustment to understand whether or not they could rely on those feeds, and what the standard of reliance on that information was.

Narrator: Ideally, legal reviews should occur early in RAS-AI design and development, to provide an opportunity to secure more humanitarian and ethical outcomes.

Damian Copeland is a legal practitioner with expertise in Article 36 reviews of weapons, specifically weapons and systems enhanced by AI. He has over thirty years of military service.

Damian Copeland: Defence complies with Australia's legal obligations. These represented in the body of law known as International Humanitarian Law, or the Laws of Armed Conflict, LOAC. That places obligations on individuals, whether they be the commanders or the operators of artificial intelligence in armed conflict and makes them responsible for the lawful use of artificial intelligence.

There are policy frameworks that are relevant within defence. Defence recently published their doctrine on ethics, which is an important framework that applies. And of course, Defence has policy that gives effect to the legal obligation to conduct what's known as an Article 36 review of new weapons, means, and methods of warfare.

That obligation essentially requires defence to ensure before they employ any new weapon, means, or method of warfare; that it is lawful in relation to Australia's legal obligations.

The development and introduction of RASAI in defence is at an increasingly rapid pace. And there is a need for defence to consider, whether there is sufficient policy framework to properly enable and regulate and govern the development of the use of RASAI.

The Australian government has a national policy on ethical AI. There are other relevant guidelines that exist both at the federal and state level. But the question is whether defence needs our policy that specifically addresses the legal, the ethical, and the safety issues that are related to the employment of AI in a military setting.

Narrator: Full testing and evaluation at the early stages of RAS-AI development and throughout the system's life cycle, is a vital part of the legal review.

The 2021 Perry Group Paper "Striking Blind" provides an example of the necessity of legal review by depicting a fictional scenario, which describes the risks of rushing technology into use.

SQNLDR Sean Hamilton: I'm Sean Hamilton. I'm a pilot with experience on Super Hornets classic coordinates, and a little bit of time flying remote piloted aircraft. Our story was Striking Blind, and our sponsors from Army Headquarters wanted us to explore the impact of trust in autonomous systems.

The Striking Blind story is about the ADF putting an Artificial Intelligence system called Mandela into service. Mandela think of a room-size, Artificial Intelligence System.

Theoretically, it's a decision AI. It helps commanders make targeting decisions. Our story was developed by the United States to fight a high-end war against China. Because the high-end conflict decisions were so rapid and complicated that humans weren't able to make them appropriately. So, they fielded an AI.

Australia bought that AI under pressure from the United States, with the U.S. highly recommending that Australia purchase the AI, and put into service to facilitate integration and interoperability in coalitions.

In addition to that, Australia was doing modelling, looking at their likelihood of success in a war against China, and realized that there would be at a significant disadvantage if they didn’t purchase the AI. So, there was a lot of pressure to go ahead and filled it.

Fast forward, our story is not set in a high end war. It's set in a low-tech, separatist conflict in the Philippines, where we're assisting the Filipino government re-establish control over areas of territory. Mandela AI is there with the Ground Force Commander, and it's helping identify targets using all-source intelligence so that we can engage threats to friendly forces more effectively.

The story starts with a crew of a Super Hornet Airborne over the Philippines. They are escorting a Filipino Special Forces Patrol that's down on the ground. They are accompanied with some Loyal Wingman drones made by Boeing. They're orbiting overhead the Special Forces Patrol, and they start identifying targets via the drones.

They're unsure whether they are hostile initially, but they are closing with the friendly patrol. They are deemed hostile then by Mandela. From the Super Hornet crew's perspective, they're looking at a display of a bunch of trucks that are yield of with a whole lot of contributors, and as Mandela identifies them as hostile, they're flipping red on their scope.

Now, the Mandela or AI could be set to just automatically engage those targets, or directly engagement of those targets. But in this particular conflict, to mitigate the risk of having an autonomous system, employing firepower without a human in the loop.

The Mandela AI is just recommending targeting suggestions to the commander, and then the commander on the ground is target. Engagement Authority is ordering strike. So, he goes ahead and order the strike. Thirty-two enemy fighters have destroyed on the ground, and the Special Forces Patrol is protected.

That engagement is a success. The fighters are then re-rolled to a second engagement. Bridge of seven kilometers to the south, where our sensors have picked up seven technicals on a bridge, which Mandela then deems hostile because they're an imminent threat to the Special Forces Patrol. The AI thinks that they are about to step off and commenced an engagement. The commander gets that decision. They attempt to get eyes on those vehicles, but due to low cloud, they can't.

So, ultimately, Commander weighs the risks and accepts the Mandela engagement recommendation, based on it, being an imminent threat to the friendlies[?]. Weapons are dropped through the cloud and at the time successfully destroyed the technical. The story fast-forwards for years. It has now come out that that engagement was a fall-out, and it was one technical on a bridge. That was a checkpoint, stopping civilians fleeing a village. And the strike killed 92 people and injured another hundred or so.

Narrator: In part, the catastrophic error that occurs in Striking Blind is the result of an inability to access the full bounds of the learning algorithm that informs the Mandela AI system.

MAJ Rebecca Marlow: My name is Bec Marlow. I just finished Staff College as part of the Perry Group option.

Understanding what the technology is inside, so what the programming has been.

That was one of the key things. What we're drawing that from our story is that we didn't understand what the programming was and what the learning algorithm was for this black box system. And that because we didn't own the IP, and because of IP issues that we were unable to fully understand and appreciate. And because we had not conducted the full testing evaluation process and had been unable to test edge cases; an edge case scenarios that would potentially affect the employment of the system. We were unable to identify what the issues were potentially, because it just hadn't gone through the full process.

Narrator: An edge case is a problem or situation that happens at extreme operating parameters. It can be expected or unexpected. In the case of Striking Blind, the edge case affected the accuracy and trustworthiness of the RAS-AI.

MAJ Marlow: The system is being spoofed by the enemy. They were sending false signals and they had cut off all other network feeds to us as well. The biases of the original U.S. programmers that they would always have that access to the network. So, there are assumptions and biases that affect the learning algorithm for them, the system, and therefore wasn't prepared for that to occur.

Narrator: The Striking Blind paper recommends that Defence take a whole of system approach to certification and validation.

SQNDLDR Hamilton: How does the legal frameworks interact with the introduction and Mandela AI into the scenario? We didn't specifically address that point as part of the Senate inquiry, but the way that it would work is that were expected to undertake an Article 36 review; which essentially just means, it just says that you were to review new weapons that are coming into service, and essentially making sure that they still adhere to the Laws of Armed Conflict, principles of distinction, proportionality, necessity, and humanity.

In our story, it's assumed that they are adhering to those principles. Then Mandela AI is assessing those Laws of Armed Conflict Principles in its decision-making. And in this particular engagement, it made a mistake.

What's our new policy do we need in the ADF to govern the employment at the AI? Who would say that we just need a policy framework to figure out who exactly is accountable, when that AI goes into service? We need to discuss what the bounds are, for when our commanders are expected to trust the AI, and when they’re expected to question the AI. That will take a lot of work.

Narrator: Australia has a number of international legal obligations, including those imposed by the Geneva Conventions, that dictate Australia’s legal responsibilities associated with the use of RAS-AI when they are used in armed conflict.

The Geneva Conventions form part of the body of law that sets limitations on the way warfare may be conducted. This body of law is called Laws of Armed Conflict (LOAC).

These laws regulate the conduct of armed conflict and impose legal obligations relevant to Australia’s design and use of RAS-AI.

The obligation to ensure Australia's use of new weapons, means and methods of warfare is consistent with Australia's legal obligations is created by Article 36 of Additional Protocol I.

In Australia, Article 36 reviews are completed by legal officers within Defence Legal's Directorate of Operations and International Law.

Ideally, Article 36 reviews consider new weapon technology in early design and development stages, which provides a chance to review, improve compliance, and secure more humanitarian outcomes.

Damian Copeland: How would Australia conduct an Article 36 review, would be informed by whether or not Australia regarded Mandela as being subject to an Article 36 review requirement. So, the first question that Australia would consider is whether Mandela as an AI decision tool, is in fact within the definition of a new weapon, means, or method of warfare as Australia understands. And this is a policy decision that Australia would make.

Now, in this case, the Mandela's may not be regarded as a weapon per se, but it may be regarded as a means of warfare, for example, that would potentially bring it into the Article 36 review obligations. And that's a matter for Australia to determine as a matter of policy.

Narrator: The Perry Group Striking Blind Paper demonstrates that robust oversight of the RAS-AI will benefit Australian Defence.

MAJ Marlow: We still had the human in the loop. So, it wasn't a fully autonomous system. It was a semi-autonomous system being employed. Under Article 36, that is one of the four targeting systems that is actually a caveat of Article 36; is that you still have a human in the loop that the robotic system cannot make the decision itself to target things. There has to be someone else, in the end, making that goal.

Damian Copeland: If Australia was to conduct an Article 36 review of the Mandela, I'd imagine that a lot more information than the scenario suggests was made available. There are a range of issues that are relevant. Mandela's functionality clearly entails decisions that are governed by the Laws of Armed Conflict. So, questions around, how were those Laws of Armed Conflict actually programmed into the artificial intelligence? What was the interpretation of the U.S. program, and how well did they understand the applications of the Laws of Armed Conflict?

These are only just the start of the information. The legal review would also be concerned with the data that was used both to train and to validate the system. The legal review will be interested whether there are biases that are present within the training data. Whether the training data itself was suitable for the environments in which Australia intends to use the system.

It would have to demonstrate that it is capable of use in accordance with Australia's legal obligations. That would mean that Australia would need to understand the legal rules that applied to its functionality. And then whether or not its consideration and

application of those rules meet the standards that Australia requires. That would require independent testing by Australia of the Mandela system. To make sure that the testing data or information that they rely upon to make an assessment is sufficient.

So, the scenario talks about the system being developed by the U.S. But the important fact from an Article 36 legal review perspective, is that the review applies the reviewing country's legal obligations.

The U.S. doesn't have the identical legal obligations, and interpretations of the law, as Australia does. On that basis, the Mandela system would have to be subject to a full review by the Australian Defence Force.

Narrator: As demonstrated in the Striking Blind story, we need to devote considerable efforts when it comes to conducting full review of RAS-AI.

To fully understand the design of RASAI and its alignment with Australia's legal and ethical obligations, Defence should work closely with partner governments and industry, to oversee the early creation and development of new systems. This close relationship should continue through-out deployment and use of the system.

SQNLDR Hamilton: Not having a clear understanding on how exactly Mandela is making decisions; that is a technical and legal problem with how we're going to buy Artificial Intelligence Systems from the United States; and have a clear understanding of how exactly they're developed, how they're taught, how they're learning on to fly, and how can we analyze their decisions and pick up any errors that happen in the course of operations.

That's going to require very close work with industry. We might need to be inside the tent when they're developing the systems and we need to have really clear arrangements for the flow of information. We need to make sure that when we purchase the Artificial Intelligence System, it's not just buying at the shop and taking it home. It's a full life cycle approach where the industry that made the system is fully embedded in the operation of the system. They're helping the ADF continue to monitor its functionality, conduct continuous tests, and making sure that it's still behaving in the way that we expect it to behave, so that we've got full disclosure on how exactly it's making decisions.

MAJ Dominic Tracey: The better ethical and lawful decisions that could be made through AI, need to be brought forth early in the pace within the learning cycle of AI system itself. Just like a person, if you give the AI a good information and good data, you can then test that person or in this case, the AI to see what decisions it makes, and then validate the decisions were correct or incorrect. And from there, I understand how your morals and ethics align with those decisions or how they may not. And more importantly, how do you change the way the AI thinks or learns; to then be more appropriate decisions at the other end, all recommendations to commanders for decisions.

Within AI being introduced into Australia into the near future. There has to be a clear line of legal requirements for commanders and end-users. And an impetus placed on

the people developing the training and technology, to make sure that they're aware of their legal liability for the future, on decisions based on the recommendations of AI.

It allows people to be invested in. The insurance that the AI is going to provide recommendations that are appropriate for our organization in a conflict, in the future.

Narrator: So, when it comes to the development of RAS-AI, what are the limitations on programming legal obligations into a machine and where will humans still be responsible?

Damian Copeland: The AI with a machine should be able to operate consistent with Australia's legal obligations, and obligations around distinction precautions in attack proportionality, not causing unnecessary suffering or superfluous injury. All of these are relevant to the legal review, but it's not that the responsibility for the compliance is delegated to the machine because that's not the case. It's the operator. It's the human who remains responsible for the use of the artificial intelligence. And so, the review is concerned with whether or not it can be used in accordance with the legal obligations. To delegate the responsibility, would potentially risk gaps in accountability and responsibility in the use of the system.

The Laws of Armed Conflict clearly written for humans to comply with. Some rules are bound to be more difficult for an artificial intelligence than it is for a human. And that's because some of the rules require distinctly human judgement. So for example, the rule of distinction in relation to civilian objects, requires that where there is doubt, as to the categorization of an object, as a civilian object, that doubt requires the presumption that it is in fact, a protected civilian object. So programming doubt into an artificial intelligence may be a very difficult thing to do.

Alternatively, distinction, which requires distinguishing between lawful and unlawful targets, may in some circumstances be eminently programmable into an artificial intelligence. But those circumstances might be quite limited, and that's where the artificial intelligence can use very clear, very discrete data to determine whether something is a military objective or a civilian object. And so for example, if something emits a particular signal or has a unique characteristic that the artificial intelligence can identify, then those type of scenarios may be more achievable than others.

Narrator: Understanding and meeting the legal requirements for RASAI are crucial when it comes to ensuring that a system complies with International Humanitarian Law and passes Article 36 weapons review.

Defence and Defence industry will have to work closely to create appropriate assurance frameworks for RASAI and ensure compliance with ADF’s requirements so that systems are lawful and ethical. Defence is responsible for the lawful and ethical use of RASAI. This responsibility includes ensuring appropriate consideration of legal requirements and ethical risks arising from the design, development, and use of RASAI's capabilities now, and in the future.

[END]

Traceability – Transcript

Narrator: In Defence, it is vital that we can explain how and why decisions are made and how and why events occurs. This transparency will reinforce the trust that people have in Defence.

In the same way, decisions and events that occur via robotics, autonomous systems, and artificial intelligence (or RAS-AI) must be transparent and explainable.

This is why Defence identified traceability as one of five facets for ethical AI. This video is part of a series that explores the facets, which are responsibility, governance, trust, law, and traceability.

This video considers traceability and RAS-AI from the perspective of Defence personnel. It uses examples from Army use of unmanned ground vehicles and a fictional scenario called “Striking Blind” to understand the importance of embedding traceability at the early stages to help us make sure we can record and audit different aspects of RAS-AI.

In the Australian Army Lavarack Barracks, Corporal Lachlan Jones is working on developing new Robotic and Autonomous Systems. As part of this work, he developed an app for a system that helps soldiers send reports.

CPL Lachlan Jones: I'm Corporal Jones. I'm a member of a support company working in 1RAR innovation space. Over the last 12 months, I've been working on photogrammetry and point cloud data that can be generated and disseminated on the tactical edge of the battlespace to enhance section commanders' level of awareness.

Narrator: Traceability is an essential aspect of his work. Traceability is concerned with tracking and reviewing the use of the RASAI by identifying and maintaining records on events and information created and used by the system. Transparent record-keeping enables traceability which is important as it allows us to address the ethical and legal aspects of the system and review events when poor outcomes happen.

CPL Jones: With RASAI, I can automatically process data. It may place a low significance on potentially important information that could change the outcome of a mission. I think trusting something that can potentially make decisions without any human input can be hard. But now through working with autonomous programs that collate and process data, I trust those to process data without any input. Whereas originally, I would constantly track them and monitor them.

Narrator: Defence is proactively considering the ethical issues that may arise with RASAI and working to make sure these technologies are used within traceable systems of control. Australian Chief Defence Scientist Tanya Monro has said this consideration of ethical issues should occur in parallel with technology development.

Professor Tanya Monro: My view is that traceability is central to the social license

we have as a Defence to use AI and thus it is essential. We may not now today have the full knowledge we will need in the future to be able to do that tracing, but I put it to you that we need to make that a focus. And the reason is the Defence organization has to work on the lessons-learned principle. It's not acceptable for us to have an outcome in conflict or even in the gray zone, where we can't understand how the data led to the information on which a commander or someone in the field made a decision.

So, traceability is core. What it does, is it focuses the mind on the interface between the human and the AI. How far does the human have to be into the loop or on the loop to give us that traceability? And how can we extract the lessons learned from these trials, exercises, and real conflicts? It's about risk. We take on risks anytime we use technology, whether that be a piece of military hardware or a software solution.

And in understanding how we ethically use force and generate force, we have to be able to go through lessons learned and improve the way we use them in the future. So traceability is king. Rendering trust requires two core elements. The first is transparency, the awareness, and understanding of how data is used by AI to provide information to decision-makers. The second is verification, the ability to be able to trace through the system, how those AI algorithms produce outcomes.

With transparency and validation, we can build layers of trust into our system.

Narrator: In the situation described by the 2021 Perry Group's Striking Blind story, the Mandela system is opaque by design because of intellectual property rights and export control. Within this story, the lack of transparency around how the algorithm learns from its environment and resulting changes in assessments of the battlespace has severe consequences.

MAJ Dominic Tracey: I'm Dominic Tracey. I'm a Logistics Ordinance Core Major with a petroleum background. And I was involved in the development of Striking Blind, which was a Perry Group paper to look at AI for the future and implementation in the battlespace.

So within Striking Blind, we looked at an autonomous system and we called it the Mandela system and it was a decision support tool within a fighter pilot or the kill chain. What we wanted to look at, was the ethics behind that and how that AI could change the battlespace for the better in the first instance by then some of the risks that may not be as obvious, particularly when it was introduced into service.

Unfortunately, for humans, it takes a lot longer to go through that process to test and evaluate an AI. And then for humans who are involved in that situation, the shorter the period they have to consider information, they'll have less information than they can filter through to make that decision. And those are the kind of areas where AI has the potential to increase our lethality on the battlefield but is where it has the greatest risk if we don't understand how it makes those decisions.

Some of the frameworks with governance that has to go around AI; firstly, understanding how it makes the decisions to start with. And then, secondly, testing

and evaluation post, any decision that is made by an AI or recommendation more importantly made by an AI before we trust the systems.

Narrator: The fictional Mandela system uses data from mobile phones, wi-fi routers, and other sources to identify who is present in the battlespace. Unfortunately, this data is corrupted by deliberate spoofing and over time the Mandela system learns to treat this fake information as an accurate representation of who is present.

MAJ Tracey: So within Striking Blind, our main protagonist Charles engages ground targets based on a feed from the AI saying, there is a legitimate target and what leads up to that is a ground commander calling for fire support. The AI system identifies a group of targets that are legitimate targets and they're engaged. What it doesn't do is translate that into a secondary target very well.

So the same pattern of behavior is identified through a secondary group, which turns out to be a civilian group. And then, the same pilot is given the same information feed, makes the same decision based on that information feed. However, it strikes an illegitimate target and causes unacceptable collateral damage.

Narrator: in the Striking Blind story, the fictional AI identifies an enemy target. However, it later emerges that the people targeted and harmed were refugees fleeing the conflict occurring in the scenario. This story illustrates why transparency in RAS-AI systems is vital.

SQNLDR Sean Hamilton: Sean Hamilton. I'm a pilot with experience on super hornets, classic hornets, and a little bit of time flying remote piloted aircraft.

It learns that spoofing was a normal part of the environment and is expected to be there, so, it ceased to question the validity of the transmissions that it was sensing and it slowly move down the learning path, which was adverse and resulted in adverse consequences. So, what we need to be able to do is to be able to continuously monitor the AI to figure out how exactly it's making decisions so that we can detect these sorts of issues coming up and we can maintain oversight of the learning algorithms.

Narrator: Monitoring RAS-AI will require technical skill, at least in the near-term. Partnerships with technology developers and training for Commanders and operators will be an important pathway to enabling this aspect of traceability.

SQNLDR Hamilton: Now, these learning algorithms, they're going to be very complicated, it's going to require a strong partnership with the industry. So you would imagine the developer of their AI, so in this particular case, we call it North Well Atomics would need to be embedded with the AI helping us maintain oversight of it.

Narrator: Users of AI will need to determine how to provide these levels of explanation using international standards and best practices. Traceability also includes lookup information for commanders and operators. They will need to be able to access and use the information on the RASAI in question and through this understand the technology's capabilities and limitations.

The Australian Army has taken a learn by doing approach to RASAI, with some systems piloted in trials. Major Chris Hall has worked on trials that use robotic vehicles for the resupply of combat teams and unmanned aerial systems for reconnaissance and resupply. He is exploring the best ways to prepare future army leaders to work alongside machines and human-machine, teaming more broadly.

MAJ Hall: Commanders will need a general understanding of the capabilities and the likely actions of a RASAI system without understanding all the technical aspects behind it.

At a user's level, something like an activity log might be useful. At a Commander's level, something like the aggregated data from all the systems available in their force might be useful. But we can often explain these things in non-technical terms. If you look at a book, like I Robot, it does an excellent job of explaining autonomous decision-making and where that can go wrong without pages of code. So there are other tools out there that we can use to teach commanders and users how a RASAI system will likely act.

Narrator: Corporal Jones has experienced responding to feedback from an operator's perspective. This feedback process is important if Army is to build technology, that is as useful as possible for them.

CPL Jones: Soldiers need a comprehensive training package to enable them to operate RASAI in complex and stressful situations. The limits can be understood and communicated early on in soldiers' careers when they first joined the army and all their life through their employment training as well as continuation training within the workspace.

Narrator: Providing this information is not just about helping our Commanders and operators to use RAS-AI to their best effect. It also ensures that we are complying with our legal obligations.

Damian Copeland is a Senior Research Fellow in the University of Queensland Law and Future of War group. Damian’s research focuses on the application of export control, arms trade and sanctions regimes relevant to the export and brokering of trusted autonomous military systems and associated technology.

Damian Copeland: My name's Damian Copeland. I'm a research fellow at the University of Queensland and part of the Law and the Future of War research team. Our role is to investigate how the law both enables and constrains the use of autonomy in Defence.

So it's important in terms of information required for a commander to have sufficient information to use AI to achieve a military objective. Now, part of that is that the commander is the responsible person, not the AI. So the law applies and holds a commander accountable and others for the lawful use of the AI, which is one of the many tools that may be available to combat or in any particular situation.

Narrator: RAS-AI also present an opportunity to improve existing processes for collecting information during conflict for assessment or review.

Dr Lauren Sanders: My name is Lauren Sanders and I am a Doctor of International Criminal Law. I've spent 20 years in the military as both a signals officer and a legal officer. And my area of expertise is largely operational law and the application of International Humanitarian Law to ADF operations.

I think there are many opportunities for us in developing new technologies from a Defence and an Australian National Security perspective. One is obviously, to ensure that we are aligned with our peers when it comes to our military capabilities. But the second is also to use these technologies to enhance our ability to demonstrate our compliance with international law and to provide assurance about the moral and ethical authority of the decision-making and processes that the ADF utilizes to discharge force or utilize force.

So a great example of how UAS can provide that is the use of UAS to provide battle damage assessments. The fact that we can record that the ADF can record after they just release a weapon, the impacts of that weapon in real-time to be able to identify what has happened. Assist in demonstrating what the impact of that use of force decision was.

Narrator: Traceability requires that Commanders and operators can understand RAS-AI systems and monitor their performance. Other mechanisms for traceability include record-keeping mechanisms that enable after-action review.

After-action reviews provide a process for analysing what happened, why it happened, and how it could have been done better.

MAJ Rebecca Marlow: My name is Bec Marlow, just finished Staff College. As part of that, I did the Perry Group option and we wrote a paper as part of a group module looking at some of the issues that could be faced and challenges that could be faced for future AI procurement within the Defence. So that was what we were looking at was what happens when you don't go through the complete process.

After action reviews or post-activity reports and those sort of things, recording what happened and the process that was went through, especially with a targeting system, to understand what the process was before you went through that. So we already have those types of reports available to us, so just using those, something that already exists within our system and using those to record decisions, then protects future investigations and all that sort of stuff so you understand why the decision was made in the first place and then can look at that and review, and whether or not changes need to be made to the way decisions are made or changes to the system itself, to enable better decisions to be made.

Narrator: The space of RAS-AI in Defence is constantly evolving. This constant change and development mean we have to be able to trace what happened and why.

MAJ Tracey: Coming into the future when it comes to recording RAS-AI recommendations, I think until there is confidence in the system that is assured through our extent frameworks or future developed frameworks, it would be

inappropriate to make decisions may be based on a recommended course of action from an AI. If that decision-making process isn't recorded, that is a question of confidence and competence in using AI which Australia will not be at in the near future.

Narrator: A record-keeping mechanism should be embedded within RASAI to enable thorough after-action reviews.

MAJ Tracey: So, when it comes to decision-making and recording of those decisions, there has to be a recording mechanism of the AI itself, which means that Australia needs access to the IP of that AI to understand how it made the decision and then to test what data was used for that in making that decision.

Narrator: To ensure success in the development and use of RASAI, the Defence will need to investigate ways of enabling useful record-keeping. Record-keeping can represent the technology involved, the chain of events, and the humans and machines that were part of the decision-making.

Through record-keeping, Defence will be able to rewind the decision-making process to understand what has occurred during the use of RASAI and what lessons might be learned. This information should be accessible and understandable for different groups of stakeholders. This may mean different levels of information are provided for experts and non-experts.

There are technical hurdles involved when it comes to recording data. However, at the Striking Blind story demonstrates, overcoming these hurdles is vital.

MAJ Tracey: So, how can we make sure that when we employ an AI that appropriately records information so that after the event, we can go back and figure out exactly how it made a decision. So all of the data that the AI is sensing is going to need to be recorded and then data on how exactly it's making decisions, that's going to be technical hurdles with that.

So theoretically with this particular AI, even if we can see exactly how it's making its decisions, the way that it's making its decisions because it's all based on machine learning might not make sense to humans. So, the information that it's using to figure out that this target is hostile, might not be in accordance to what you believe to be common sense.

So for example, in engaging these seven technicals at the bridge, maybe the AI is sucking the data from the internet. And it seemed that people are posting fewer cat videos on YouTube and it is associated with having fewer cat videos with insurgent activity in this particular area and therefore, it's making that decision.

Narrator: Developers of RAS-AI must seriously consider how to best enable traceability via record-keeping in a way that supports Commanders and operators and does not introduce additional burdens.

MAJ Chris Hall: Currently, we only keep detailed records of values of RASAI systems where something has gone wrong but they're used currently in a logistics

and resupply and reconnaissance role. As they gain more autonomy and lethality, more detailed record-keeping may be required, but that needs to be an automated and electronic process.

It can't become an administrative burden that must be manually undertaken in battle, or it simply will not happen in war. There might be a utility for commanders in the aggregated data that can be taken from all these systems, but we need to apply some common sense as to how much data a commander can take in at any given time.

More information available doesn't mean we have more human capacity. The types of records to be kept might depend on the system. A user-level activity log could be useful, commanders aggregated data could have utility, and for particularly complex lethal systems, they might be more detailed records than that.

However, we should be cautious of overburdening ourselves. And most of the current weapons and mostly our current vehicles don't have that black box capability. So we already accept that information is going to be lost in war. And then, you throw attrition over the top of that with many of these RASAI systems likely to be considered attritable.

We're not going to get all that information back off the battlefield, so we can't make that a requirement. At times, we may have a moral imperative to employ RASAI systems even where their decisions will not be fully recorded if that will prevent greater consequences such as losing a battle or taking human casualties. Uncertainty has always been part of the enduring nature of war and information has always been lost in battle and we can't expect that to change through the introduction of RASAI systems.

Narrator: Returning to the Striking Blind story, we can see that part of what was missing in the Mandela system as described in the story was transparency.

MAJ Marlow: The limits of the system was a difficult one because it still hadn’t gone through the full process, so with a lot of our systems that we have, like any systems within Defence, there’s always a training progression for different types of users, so what the user of the system and what the commander of the system learns about the system is different, so you’ve got to have those training courses in place, and with the Mandela system we still hadn’t gone through that whole process. So we were still in the middle of doing the testing and evaluation process when it was asked to be introduced in to service. So in this instance the Commander didn’t have the full picture of what the limits and limitations of the system actually were in this instance. And that’s something that is important especially with a targeting system to understand what those limitations are.

Narrator: Decisions are made using RASAI in Defence should be traceable and explainable. This means that the technology's data training, theoretical underpinning, decision-making models, and actions should be recordable and auditable.

Information on events and decisions related to RAS-AI should be accessible and useful for multiple types of stakeholder groups, such as users of the systems, safety certifiers, lawyers, accident investigators, and other non-expert publics.

These explanations should help us ‘rewind’ decision processes and events to understand what occurs and what lessons can be learnt for the future.

Importantly, measures taken to achieve this traceability should support commanders and operators to achieve effective, lawful, and ethical outcomes.

[END]

Animations

Responsibility (Animation) – Transcript

Robotics, autonomous systems, and artificial intelligence (RAS-AI) will deliver game-changing capabilities for Australia's defence.

AI can improve aspects of human-decision-making, but it must be clear that humans are legally and morally responsible for decisions or actions that use AI.

Defence identified responsibility as one of five key facets of ethical use of RAS-AI because humans are responsible for the technologies they employ.

The ADF Doctrine on Military Leadership states: "Ethical leadership is the single most important factor in ensuring the legitimacy of our operations and the support of the Australian people." And this is reinforced in the Doctrine in Military Ethics.

To use RAS-AI systems responsibly and act effectively and ethically, Defence must understand how an AI system has been developed; how it behaves; and how to use it, including the potential consequences of its use.

Education is critical to enable a Commander to enact their responsibilities, particularly in the use of combat systems.

Decisions made using AI must be captured using accountability frameworks that designate ethical and legal responsibility.

This is especially important for decisions relating to the use of force, which require clear lines of human responsibility and accountability attributable to relevant commanders and weapon operators.

Human-machine teams must behave in accordance with ethical frameworks to empower human agency, enhance action, and ensure moral responsibility.

A human-centred approach will help ensure that human beings are ultimately responsible for decisions made supported by AI.

[END]

Governance (Animation) – Transcript

Robotics, autonomous systems, and artificial intelligence (RAS-AI) will be used in many aspects of Defence in Australia). RAS-AI must operate in accordance with Australian governance, including: laws, principles, doctrine, regulation and best practice frameworks.

The ADF Doctrine on Military Ethics says that commanders must be clear on the legal obligations that govern their operations.

Similarly, personnel must understand national and international legal requirements before deployment.

That means that - as in all other aspects of Defence - there must be frameworks and tools that govern RAS-AI operations. The introduction of RAS-AI can change how decisions are made and therefore what checks and balances are required to ensure operations are lawful and ethical when using these technologies.

Defence has identified governance as one of five key facets of ethical use of these systems.

Critical for governance is understanding both the context in which technology will be used and how it will be controlled.

The appropriate level of human control may vary depending on the parameters of the system and the operational environment within which the system is deployed.

Control will be exercised over the lifecycle of a system through legal, policy, technical, and professional mechanisms.

These mechanisms include: legal obligations, ethics policies and standards, cultural values and norms, operational guidelines, safety manuals, tests, evaluation verification and validation procedures, and after action reviews.

These tools help ensure that human-machine collaboration is the best it can be. Governance mechanisms are a safeguard against poor decision-making by both human and machine. Poor decisions occur as a result of a range of reasons, including inefficient or ineffective teaming, cognitive overload, attention deficit, loss of situational awareness, automation bias or mistrust of the system.

Different forms of governance must work together to ensure the safe and effective use of these technologies in Defence in their context of deployment.

[END]

Trust (Animation) – Transcript

To be used effectively in Defence, robotics, autonomous systems, and artificial intelligence (RAS-AI) must be trustworthy and trusted to do their job as intended.

The ADF Doctrine on Military Ethics says that trust is essential for people working in Defence because they are trusted to do things that are otherwise restricted in society.

Defence has identified trust as one of five key facets of ethical use of RAS-AI in Defence.

Trustworthy AI works reliably in accordance with its intended purpose. Trust can be built through investment in best practice AI methods, education and training, rigorous test & evaluation, and familiarity with the technology achieved through its utilisation in exercises and operations.

Trust needs to be built and then reinforced between multiple parties. This includes humans; humans and machines; and machines and machines.

Trust requires competence and integrity from all actors. AI developers must build trust through demonstration of Defence values and abidance with military ethics in the operation of their technologies.

Machines and humans need to know what they do well, as well as their limits; and to communicate these to each other.

Human-AI systems in Defence need to be trusted by users and operators, by commanders and support staff, and by the military, government, and civilians.

Operators will hold multiple levels of trust in the systems they are using depending on which aspect of trust is under scrutiny. Trust may change over context and time and must be proportionate to the risks.

Finally, to be trusted, AI systems need to be safe and secure. This relies on a safe and secure supply chain.

[END]

Law (Animation) – Transcript

The ADF Doctrine on Military Ethics, says that acting in accordance with the law, is the baseline standard for ethical behaviour.

That means that when robotics, autonomous systems, and artificial intelligence (RAS-AI) are brought into service in Defence, they must be usable in a manner that complies with Australia's legal obligations.

Defence has identified law as one of the five facets of ethical RAS-AI for this reason. During the design and development of AI, creators must understand the specific legal obligations and ethical considerations surrounding the use of their technology.

Australia is bound by international law, including international humanitarian law.

Accordingly, ADF operations must be conducted in compliance with Australia's legal obligations and Government directions as reflected in rules of engagement.

All new ADF weapons, means and methods of warfare must pass an Article 36 weapon review before they can be used in armed conflict. Article 36 states that:

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

Article 36 reviews consider technology throughout design and development; providing an opportunity to review and improve compliance and secure more humanitarian outcomes.

The Method for Ethical AI in Defence supports Article 36 review with the Legal and Ethical Assurance Program Plan or LEAPP tool.

The LEAPP facilitates a trusted conversation between AI creators, users, and Defence, accelerating innovation and enhancing potential compliance with international humanitarian law.

Designing systems with these laws in mind will support greater legal and ethical outcomes by design and increase the likelihood that new technologies are adopted and deployable by Defence.

[END]

Traceability (Animation) – Transcript

There are legislative requirements for Defence to record its decision-making.

The ADF Doctrine on Military Ethics notes that it is important to explain both how and why decisions are made. That's because the ability to explain makes it easier to assess the legitimacy and accountability of the decision. And it reinforces the trust that other people place in the Defence Force.

In the same way, decisions made using robotics, autonomous systems and artificial intelligence (RAS-AI) in Defence need to be traceable and explainable.

Defence has identified traceability as one of five facets of ethical RAS-AI.

No matter how AI is employed by Defence, its data, algorithms, models, theoretical underpinning, decision-making frameworks, and actions must be recorded and auditable. Data must be managed in accordance with Defence Data policies including abidance with security, privacy and provenance requirements. Algorithm training processes including the modification of algorithms and models must also be documented and recoverable.

Records can represent the systems involved, the causal chain of events, and the humans and AI that are part of decisions.

This information should be accessible and understandable for different groups of stakeholders. Transparency may require different types and levels of information for expert and non-expert stakeholders. Creators and users of AI will need to determine how to provide these levels of explanation using international standards and best practice.

When decisions lead to expected outcomes or positive outcomes, the factors that lead to those decisions may avoid scrutiny.

However, when negative outcomes occur, users of systems will need to be able to 'rewind' the decision process to understand what occurred and what lessons might be learned. This also will inform Defence inquiries into operational incidents.

Requiring traceability of decisions made by humans and RAS-AI ensures that humans remain responsible for technologies they develop and deploy; and remain accountable for their use.

[END]