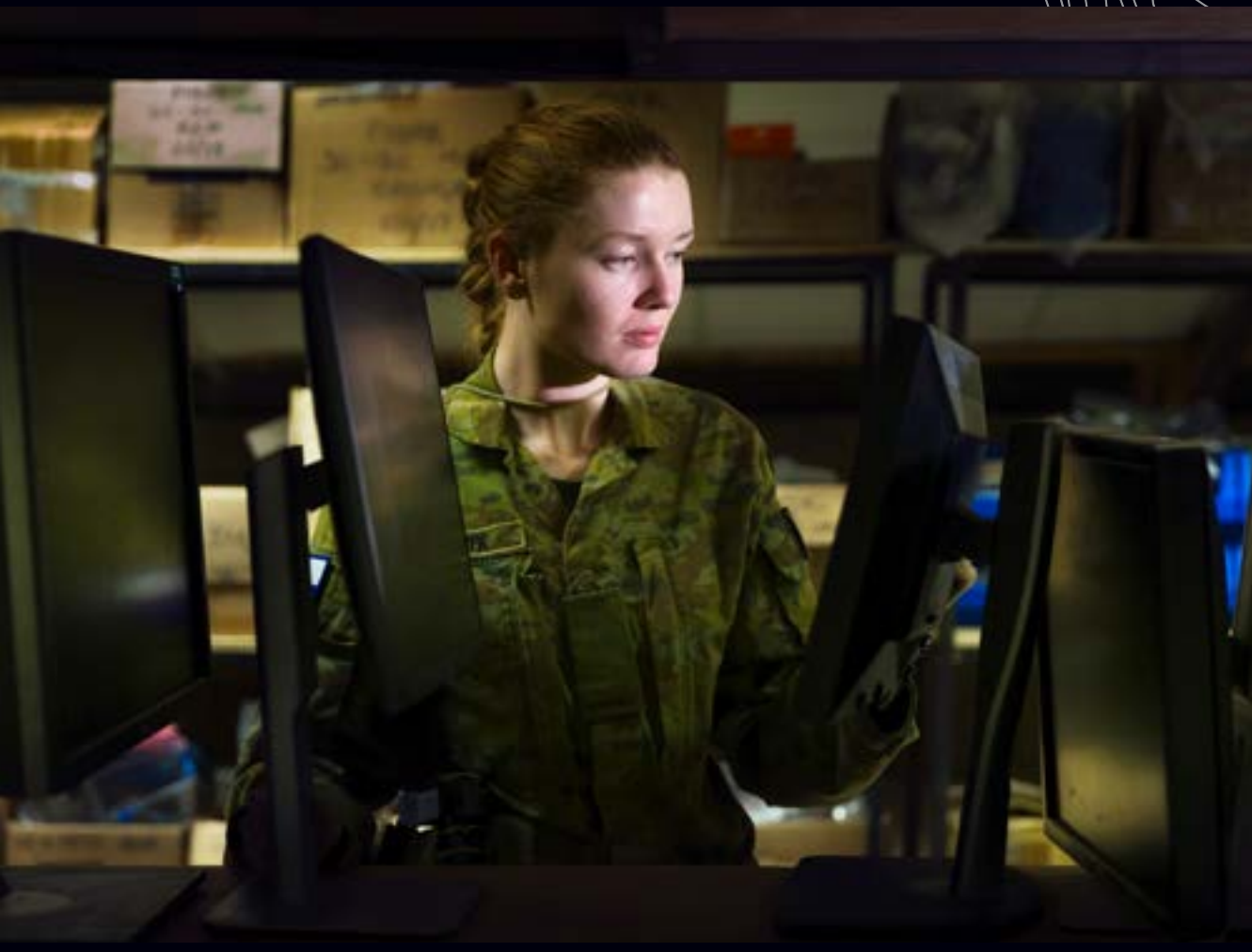
THE INFLUENCE OF VISION AI ON ETHICAL DECISION-MAKING IN MILITARY CONTEXTS

CHRISTINE BOSHUIJZEN-VAN BURKEN, DEANE-PETER BAKER, NED DOBOS, MILAD GHASRIKHOUZANI, ERANDI HENE KANKANAMGE, TWAN HUYBERS, OLEKSANDRA MOLLOY, JO PLESTED, ABI VEHDA

AUSTRALIAN ARMY RESEARCH CENTRE
/ Australian Army Occasional Paper No. 42



THE INFLUENCE OF VISION AI ON ETHICAL DECISION-MAKING IN MILITARY CONTEXTS

CHRISTINE BOSHUIJZEN-VAN BURKEN, DEANE-PETER BAKER, NED DOBOS, MILAD GHASRIKHOUZANI, ERANDI HENE KANKANAMGE, TWAN HUYBERS, OLEKSANDRA MOLLOY, JO PLESTED, ABI VEHDA

/ Australian Army Occasional Paper No. 42

ISSN (Print) 2653-0406

ISSN (Digital) 2653-0414

DOI: 10.61451/10.61451/2675163

All enquiries regarding this publication should be forwarded to the Director of the Australian Army Research Centre.

To learn about the work of the Australian Army Research Centre visit https://researchcentre.army.gov.au

Cover image: Australian Army soldier inspects computer monitors at the Taji Military Complex, Iraqi. Source: Defence Image Gallery. Photographer: CPL Oliver Carter.

/ CONTENTS

/ ABSTRACT

The use of artificial intelligence (AI) decision support tools is becoming increasingly common in military settings. The Australian Defence Force (ADF) is actively implementing human–machine teams into its operations. In parallel, it is also supporting research into the use of AI to improve efficiency and decision-making, to lower risks to combat personnel and, ultimately, to increase combat power. This occasional paper presents the findings of a research project that relates specifically to so-called 'vision AI'. Vision AI detects, tracks and labels items on a screen. While it is assumed that these labels will minimise the likelihood of error and improve ethical decision-making on the battlefield, there is little empirical evidence to support this assumption.

In November 2023, The University of New South Wales in Canberra established the Military Ethics Research Lab and Innovation Network (MERLIN), a dedicated research entity comprising a global network of researchers committed to identifying and mitigating ethical risks for military personnel. In 2024, MERLIN researchers conducted a series of experiments involving simulated battlefield scenarios. Participants were presented with choice situations—to shoot or not to shoot—in two different conditions, either with the assistance of vision AI or without it. The aim of these experiments was to determine what difference, if any, vision AI would make to the participants' decisions.

The experiments show that—at least in relatively straightforward and uncomplicated battlefield situations, where human users/operators feel confident in their own direct sense perceptions—vision AI has no significant positive or negative influence on the choices made. This outcome suggests that some of the ethical disquiet surrounding the introduction of vision AI to the battlefield is premature. The research also suggests that some of the optimism about the potential for vision AI to improve ethical decision-making in battle may also be unfounded. These preliminary findings contribute both to the ADF's ongoing efforts to modernise its capabilities and to its understanding of when and under what conditions AI-based decision support tools can best assist ADF personnel to achieve their directed objectives. Continued research will better ensure the ethical use of AI decision- support tools by the ADF and will minimise harm to human users/operators.

/ 1. INTRODUCTION

It is becoming increasingly common for industry and government to use artificial intelligence (AI) decision support tools to analyse data, identify patterns and suggest actions based on the evidence considered. These tools are used in health care (diagnostics), finance (prediction of market trends), cybersecurity (threat identification and response) and military operations.[1] The international humanitarian community considers AI to be one of the most pressing humanitarian and legal issues in modern warfare[2] and, as such, monitors and reports its use in contemporary conflicts.[3] The International Committee of the Red Cross (ICRC), for example, has detailed its views on military AI in the 2024 Challenges Report.[4] The Australian Defence Force (ADF) is implementing human–machine teams into its operations, with a major equipment acquisition program currently underway. The ADF also supports research into the use of AI to improve efficiency and decision-making, to lower risks to combat personnel and, ultimately, to increase combat power.[5] Its AI systems, along with all military-related systems, must adhere to international humanitarian law (IHL).[6]

The assumption underpinning these developments is that AI decision-making tools have the potential to improve overall human decision-making in terms of speed, accuracy and effectiveness. One kind of AI decision support-tool that is seeing increased use in military settings is vision AI, which detects, tracks and labels items on a screen. Provided the object exists in the system's digital library, vision AI classifies it into a category (for example, 'building', 'vehicle' or 'person') and labels it accordingly. It is assumed that these labels will minimise the likelihood of error and improve ethical decision-making. However, while the academic community is actively considering the ethics of defence-related AI,[7] there is as yet no evidence that vision AI does, in fact, have any such effect. This is partially because vision AI systems are novel, their use is currently sparse and there has been little empirical testing. Despite this, vision AI systems are likely to become a standard feature in military conflicts. For example, autonomous drones, which can be used to track and engage enemies without human interaction, have played a key role in the Ukrainian government's war strategy. Ukraine is also using facial recognition technology to identify potential war crimes perpetrators.[8]

Our project's aim is to empirically test the assumption that vision AI systems minimise error and improve ethical decision-making. Through our research and practical experiments, we hope to better understand vision AI's potential for positively or negatively affecting ethical decision-making in military settings. To this end, we designed and administered a series of experiments involving simulated battlefield scenarios. Participants were presented with choice situations—to shoot or not to shoot—in two different conditions,

either with the assistance of vision AI or without it. Our fundamental research question was: 'What difference, if any, would vision AI actually make on these decisions?' In a previously conducted pilot study,[9] we explored possible influences of battlefield vision AI on moral decision-making. One of the recommendations from the pilot study was to conduct more rigorous and controlled experiments, with greater care taken to neutralise confounding variables, to more reliably determine the effect of AI-supported decision-making in battlefield settings. The chosen experimental approach for this follow-up study was therefore one of *choice modelling*, including a qualitative research component to the choice experiment. The process also involved administering an open-ended questionnaire to participants immediately after the experiment.

The interdisciplinary team that conducted this research consisted of two military ethicists, one ethicist of technology, one AI and machine learning expert, two human–machine interaction experts, one behavioural economist and one expert in discrete choice experimentation. Subject matter experts from the Australian Defence Force Academy (ADFA) were involved throughout the process and had significant input in military scenario design.

Before discussing the experiments, this paper explores the background of our research, which is situated within the broader context of applied military ethics (Section 2). We then provide an overview of the ethical concerns surrounding the use of AI-enabled decision support tools in battle, such as the potential for over-reliance and 'moral deskilling' (Section 3). Turning next to the practical application of AI, we sketch the current state of AI use on the battlefield, focusing primarily on its use in drone operations (Section 4). We follow this with a generalised technical overview of AI decision support tools such as the commercially available Athena AI system, which we used in our experiments (Section 5).[10] In Section 6, we describe our methodology (the experiments' design, sample composition and participant demographics) and discuss the experiments themselves. We also consider the limiting factors which affected our results. In Section 7, we present our findings. Section 8 delivers discussion and recommendations. Finally, in our conclusion we briefly explore how our findings can contribute to future understandings of AI use and the evolution of defence force training, and make recommendations.

2. APPLIED MILITARY BEHAVIOURAL ETHICS RESEARCH

This research project was conducted against the background of an age-old challenge: how to ensure the noble conduct of soldiers in war. At the same time, the conduct of soldiers in war cannot be seen apart from the technologies they interact both through and with. Technological systems are now ubiquitous in our societies as well as in military practice, and we must pose the question: How does the use of technology relate to military ethics? Before answering, we must first elaborate on military ethics itself.

'Military ethics' is not, as many quip when first encountering the idea, an oxymoron.[11] Of all professions, the military arguably boasts the oldest and richest ethical tradition, one stretching back at least 1,000 years in the West and with strong parallel strands of thought in other cultures.[12] The historical 'just war tradition' (also sometimes referred to as 'just war theory') addresses the resort to war at the level of the state (*jus ad bellum*, literally the 'right to war'), as well as the use of force in the conduct of war (*jus in bello*, 'law in war'). More recent developments have included the ethical framework of *jus ex bello* ('law out of war'), which guides belligerents on the conditions under which there is an imperative to withdraw from hostilities; the *jus ad vim* framework ('law regarding force'), which sets the conditions for the ethical use of military force below the threshold of war; and *jus post bellum* ('law after war'), which guides belligerents on their ethical responsibilities in the aftermath of war.[13] The core of the tradition, however, remains the principles of *jus ad bellum* and *jus in bello*, which are enshrined in IHL.[14]

Jus ad bellum requires that states (i.e., sovereign entities recognised as states under international law) must meet six criteria if the resort to war is to be just. First is the requirement for a just cause—war is allowed only where appropriate reasons are in place (i.e., where it is justified). In today's environment, only one cause is considered uncontroversial: self-defence or resistance against aggression. This concession also extends to the defence of others—i.e., coming to the aid of other states or groups who are being attacked (the states do not need to be part of a formal coalition or to have specific treaties in place). The second principle of *jus ad bellum* is that of legitimate authority (some scholars add the requirement of 'public declaration' to this principle).[15] Specifically, only heads of state have the right to take their state to war, a principle designed to minimise the unnecessary and destructive proliferation of wars, unlike in the past when any noble who could raise an army felt entitled to declare war on their enemies. Third is the requirement of 'right intention': belligerents are justified in waging war only

if their intention aligns with their just cause. Using a just cause as a pretext to pursue unrelated objectives is thereby excluded.[16] These first three principles are deontological in nature; i.e., they are based on the inherent rightness or wrongness of actions rather than on their consequences. They are also considered morally binding fixed principles. The last three principles of *jus ad bello* are prudential or consequence focused. The fourth principle provides that, even if the first three principles are satisfied, the war is not justified unless it is a proportional response to the harm that provides the just cause. The fifth principle specifies that action must be undertaken as a last resort. The sixth states that war must only be initiated where there is a meaningful likelihood of success.

In an ideal world, war would not result in any harm to non-combatants. The reality, however, is that even in an era of precision-targeted munitions, 'immaculate warfare'—where precision is perfect, only military targets are attacked and collateral damage is minimal—remains beyond our grasp. *Jus in bello* takes this messy reality into account by establishing four principles to guide combatants in their use of force during war. These aim to minimise harm and suffering by considering the potential outcomes of military action.

The first principle of *jus in bello* is that of discrimination (referred to, in the legal discourse, as the principle of distinction). The other three principles pivot around this. The principle of discrimination focuses on who may or may not be deliberately targeted with violence. Often overlooked is the fact that this principle is first and foremost a permission—military personnel may intentionally target enemy military capabilities and kill enemy combatants. In so doing, the just war tradition posits, they do no moral wrong. The counterpart of this principle is that combatants may not intentionally target non-combatants.

The second principle of *jus in bello* accommodates the reality of civilian or non-combatant harm. The principle of proportionality requires that if combatants foresee that engaging legitimate military targets (as defined by the principle of discrimination) will result in harm to non-combatants or non-military property or infrastructure, they must weigh up whether this 'collateral damage' outweighs the value of achieving the military objective in question. The third principle, that of necessity, requires that force be used only where necessary to achieve military advantage, thereby protecting even enemy combatants from unnecessary harm. The fourth is the principle of humanity, which restricts means and methods of warfare to those that have not been found by the international community to be inhumane (means *mala in se*, literally 'evil [or wrong] in itself'). The deployment of biological weapons, for example, contravenes this principle.

The fact that the military profession is guided by such a comprehensive ethical framework says nothing about whether specific military forces generally comply with it. Similarly, the use of military technology has sometimes posed a threat to the *jus ad bellum* and *jus in bello* principles, even though it was expected that these technologies would enhance potential compliance. Unfortunately, there are innumerable cases of soldiers and non-state

combatants contravening both the ethics of war and the laws of war, which are largely shaped by the principles of the just war tradition.[17] These cases make it easy to overlook the fact that, when considered overall, most combatants conduct themselves appropriately most of the time.

The Australian Government and the ADF are currently facing difficult issues that surround incidents of non-compliance with the accepted ethical framework. In 2016, following rumours and allegations of possible breaches of the Law of Armed Conflict by Australian special forces soldiers in Afghanistan, Defence commissioned the Inspector-General of the Australian Defence Force Afghanistan Inquiry.[18] The report, known generally as the Brereton Report (after the lead investigator, Justice Paul Brereton), was delivered in November 2020. Brereton and his team detail credible evidence that 25 ADF personnel were involved in the unlawful killings of 39 civilians or prisoners.[19] The material they considered, and their findings, are so confronting that much of the report is redacted. The front cover has a content warning which notes the inclusion of 'objectionable material' and offers a website link for support. The Brereton Report additionally details atrocities carried out by Australians in past conflicts, stretching from the Boer War to the Vietnam War, which also make for difficult reading.[20]

The inquiry's findings sent shockwaves through Australia's military, and society as a whole.[21] The murder of prisoners and non-combatants is clearly at odds with the ADF's expectations and those of the majority of Australians; the historical record, detailing 'uncomfortable truths' which have been ignored, is at odds with the ANZAC legend.[22] But even amid the harrowing details, Brereton emphasised that the vast majority of Australian special forces soldiers who deployed to Afghanistan conducted themselves with honour. The same is true of Australian soldiers in general.

The ADF acknowledged the allegations of grave misconduct, accepted the report's findings and set in place a comprehensive plan for systemic, organisational and cultural change.[23] In particular, the ADF has committed significant resources to reflecting on, revising and expanding its training and education efforts in the area of military ethics. This response builds on the ADF's past and existing programs; military ethics education and training has been carried out in the ADF for many years. For example, two of the contributors to this paper teach military ethics to trainee officers (Army and Air Force cadets and Navy midshipmen) at ADFA in a program that stretches back to at least 2009. Similar programs have been underway at the Australian War College and other parts of the ADF for many years. In addition, various units have implemented ad hoc military ethics training as part of their professional military education programs.

The Brereton Report has galvanised military ethics training and education in the ADF. Driven by an internal Special Operations Command initiative, a standardised package of ethics and training materials was commissioned. Its development and delivery put the

ADF among the world's leading military forces in this regard.[24] Under the direction of the Centre for Defence Leadership and Ethics (situated at the Australian Defence College), this package is delivered across the entire ADF, ensuring a level of consistency that was previously absent. Centralisation of responsibility for ethics education, along with the elevation of the centre's authority within the ADF, bodes well for the future of military ethics education and training in Australia.

Another important military ethics development within the ADF is the 2021 publication of the ADF philosophical doctrine ADF-P-0 *Military Ethics*. Written and developed by the ADF, this foundation document (which is available as an open-access e-book and audiobook) lays out the ADF's approach to military ethics and aims to guide ethics-based professional development across ranks.[25] This is one of a suite of six 'philosophical' doctrines that sit above all other ADF doctrines; the others are ADF-P-7 *Learning*, ADF-P-0 *ADF Leadership*, ADF-P-0 *Culture in the Profession of Arms*, ADF-P-0 *Command*, and ADF-P-0 *Character in the Profession of Arms*. Australia is arguably unique in having such doctrine. Not unexpectedly, there has been some debate over the content of the ethics doctrine, mostly over the general ethics decision-making model it puts forward.[26] Nonetheless, the ethics doctrine is an important step forward for the ADF.

Most military ethics scholars are, by academic specialisation and training, philosophers (as are three of the authors of this paper).[27] Consequently, the scholarly literature largely focuses on conceptual analysis of key ideas relevant to military ethics (most notably the principles embedded in the just war tradition)[28] or on arguments around the application of those ideas to a military's current or potential operations (for example, the debate over whether the employment of lethal autonomous weapons constitutes a grave violation of human dignity).[29] Put crudely, the questions posed by much of the literature focus on 'What is the right thing to do about x?' or 'What is the right thing to do in situation *y*?' These are vital questions but, as we are learning, they are not enough on their own.

The Brereton Report and other reports concerning failures in battlefield ethics (such as Dr Samantha Crompvoets's 2016 report)[30] have brought to light another important question that has been critically under-researched. In 2011, after Royal Marine Sergeant Alexander Blackman shot dead a grievously wounded Taliban prisoner, he stated to his teammates, 'Obviously, this doesn't go anywhere, fellas. I have just broke [sic] the Geneva Convention'.[31] There was no consideration here about whether the action was or was not the right thing to do—Blackman clearly knew that what he was doing was wrong. The key question in this case, and in many like it, is 'What are the factors that lead combatants to act in contravention of what they know to be ethical?'

This is the question at the heart of what Deane-Peter Baker has dubbed AMBER—applied military behavioural ethics research. Max Bazerman and Francesca Gino have defined behavioural ethics as 'the study of systematic and predictable ways in which individuals

make ethical decisions and judge the ethical decisions of others that are at odds with intuition and the benefits of the broader society'.[32] Behavioural ethics is thus, first and foremost, descriptive and empirical in nature, in contrast to the normative approach to ethics that dominates most military ethics scholarship. Research in this area is inextricably interdisciplinary and draws on the insights of moral philosophers, psychologists, sociologists, historians and others.

While there is a useful and growing body of research in behavioural ethics in general,[33] including retrospective considerations of actions,[34] very little attention has been paid to military personnel making ethically challenging decisions in recent and current military contexts. To address this gap, in November 2023 the University of New South Wales established the Military Ethics Research Lab and Innovation Network (MERLIN) as a dedicated research entity. As well as offering state-of-the-art research capabilities, MERLIN has a global network of researchers who are committed to identifying and mitigating ethical risks for military personnel.[35] The research project that is the subject of this occasional paper falls under the auspices of MERLIN. Together with experts in human performance in socio-technical systems and behavioural economics research, we put the ethical questions encountered in battlefield scenarios to the empirical test. Here, we were explicitly interested in the role of AI in behavioural ethics.

/ 3. THE ETHICAL RISK OF OVER-RELIANCE AND MORAL DESKILLING

In war, the use of AI-enabled decision support systems presents a number of ethical risks. One of the risks relates to the potential for *over-reliance*. This is where human operators exhibit excessive deference to these systems, accepting as accurate the information they produce and following their recommendations even where they conflict with the operator's own judgement, critical thinking and/or situational awareness. Mistakes can thus be made.[36] For example, a commander might accept an AI's targeting recommendation without verifying its alignment with rules of engagement or operational objectives. Some measure of trust in these decision support systems is certainly desirable for their effective use. Nevertheless, it is important to recognise that even the most advanced AI-enabled support systems currently available are not infallible—due to errors in their data or sensors, or biases in their algorithms, illegitimate targets can be mistaken for legitimate ones.[37] This being the case, 'over-trust' is a worrying phenomenon that has attracted considerable discussion.[38]

Over-reliance may occur for a number of reasons. One is the so-called 'automation bias'. This term is used in social psychology to describe the propensity of humans to place excessive trust in computerised systems and other automated aids. We may do this because we exaggerate the analytical superiority of the technologies in question, but ultimately automation bias is thought to be a function of a kind of *laziness*.[39] The human mind is sometimes described as a 'cognitive miser', hard-wired to conserve energy and, therefore, to solve problems and make decisions in the simplest and most effortless way possible.[40] Since AI-enabled decision support tools present humans with an opportunity to further reduce the cognitive costs associated with making hard choices, the potential for over-reliance and laziness is hardly surprising.

In extreme cases, automation bias leads humans to relegate themselves to a largely observatory role, an appendage to the machine or a 'rubber stamp', strongly disinclined to challenge the computer-generated outputs or to act independently of them. This can lead to two kinds of mistakes: 'commission errors' and 'omission errors'. The former occur when a human user/operator accepts—without question—the information provided by a system or follows its automated recommendations without factoring in conflicting information from other sources, including his/her own direct sense perception. The latter occur when an automated system fails to make an indication or recommendation that would otherwise lead to action, and the human user/operator fails to act because of this.

A second driver of over-reliance is what is sometimes called *moral buffering*. This is a 'motivated' or self-serving (though not necessarily self-transparent) tendency to make decisions in ways that allow us to minimise guilt or shame (both moral emotions). A human user/operator who relies on personal intuition and discernment when faced with a dilemma may sincerely believe that the decision made is the optimal one in the circumstances. But if things go wrong, it is possible the human user/operator will find it difficult to avoid pangs of conscience. On the other hand, if an operator defers to an automated decision support system and things go wrong, the sense of responsibility may be diffused or shared with other members of the organisation, including those who decided to enlist the support of the AI systems in the first place. This moral buffer can be a protective or mitigation mechanism to avoid or soothe a troubled conscience or aversive emotional states—including some that are liable to mutate into debilitating 'moral injuries'.[41] Therefore, the human user/operator may show a strong preference for diffused responsibility even where it is not warranted or is not conducive to fulfilling the mission.

Ethical decision-making is a skill. One must be able to recognise all the ethically salient features of a situation, to weigh them up and deliberate on them carefully, to integrate them into a well-justified 'all things considered' judgement on the right thing to do, and to conform one's actions to that judgement. Just as muscles atrophy if they are not used, the skill of ethical decision-making may gradually erode in those deprived of the opportunity to exercise it regularly. Habits of reliance on systems that abstract or simplify moral considerations can dull a person's sensitivity to ethical complexities. Over-reliance on technology can lead human users/operators to make flawed decisions and to produce harmful (or at least suboptimal) outcomes, but that is not all. If these individuals *repeatedly* outsource their decision-making to AI-enabled systems, there is also a danger that, over time, they will lose their capacity for moral reflection, agency and complex decision-making. Affected human users/operators will then have no choice but to over-rely on the technology, since their capacity for independent ethical judgement will have become too compromised. This is the danger of *moral deskilling*.[42]

Moral deskilling is a concern because, inevitably, human users/operators at all levels will find themselves faced with problems for which automated systems cannot offer solutions. These situations will arise not only during the average ADF employee's military career but also after they return to civilian life. If human users/operators do not have the opportunity to use their own decision-making capacities or to draw on their own moral frameworks for nuanced and sophisticated reasoning, these personal resources may be significantly compromised. Indeed, they may lose them entirely. If we want people to retain their ability to confront ethical challenges in service and in civilian life, we need to be mindful of the danger of moral deskilling.[43]

/ 4. CURRENT USE OF BATTLEFIELD AI

The development of unmanned aerial vehicles (UAVs—drones) has increasingly focused on integrating them with AI visual systems for terrain mapping to aid navigation and target identification. More complex programs enable interconnectability: the UAVs can operate in interconnected 'swarms'.[44] Drone-based vision AI systems have been evident in recent conflicts. In this regard, Russia's war in Ukraine is possibly the first international conflict in which both sides have actively developed and used AI for military purposes. The emerging areas of AI use in this war are geospatial intelligence for object recognition to detect the identity of invading Russian troops, real-time analysis of unencrypted Russian radio transmissions, and imagery of Ukraine terrain.[45] While debates remain about the current effectiveness and future potential of AI-enabled drones,[46] there are several key areas where AI promises to make a significant impact. These include terminal guidance, visual navigation, target detection, and swarming.

In 2023, Ukraine installed AI target identification systems on drones.[47] This development facilitated long-range drone strikes on military facilities and oil refineries 2,000 kilometres inside Russia. While such systems cannot yet outperform or replace a well-trained warfighter in performing the same task, data is becoming increasingly available and is being used to train AI systems to improve object identification. For example, by centralising footage captured by drones, the Avengers AI platform, which has been developed by the Ukraine Ministry of Defence's Innovation Centre, is now able to identify 12,000 pieces of Russian equipment a week.[48]

/ 5. TECHNICAL OVERVIEW

Object detection is a subset of the computer vision discipline that enables a computer (and thus the human user/operator) to detect the type and location of objects in an image.[49] This type of detection capability has been vastly improved by the advent of machine learning and new, massively parallel computing hardware (i.e., systems that use specialised processors to simultaneously perform calculations or handle large-scale, complex data processing tasks). The rapid advancement in the computer processing power of drone hardware has already transformed the battlefield:[50] drones now have the hardware capabilities to achieve AI inference for interpreting visual data rather than just sensing it. Efficient object detection models such as YOLO-LITE[51] and Mobilenet[52] can be trained on inexpensive computer hardware such as graphics processing units, which were originally produced by NVIDIA to accelerate computer graphics for the consumer market. Once trained, these neural networks can be deployed for object detection inference on much smaller and less expensive pieces of computer hardware, including some that are increasingly being integrated in drones.[53]

As drone-based AI applications continue to improve, there will be an exponential increase in the need for computer processing power to identify and locate objects of interest (such as tanks, artillery and anti-air defences) without the assistance of a human operator. Such a capability will substantially shorten the time required for each cycle of the 'observe, orient, decide, act' (OODA) loop, a rapid decision-making process that is critical to the effective achievement of battlefield outcomes.[54] Before considering how modern computer vision models might assist in achieve this, it is useful to explore the concept of machine learning.

The definition of machine learning was formulated in 1997 by Tom M Mitchell:

> A computer program is said to learn from experience *E* with respect to some class of tasks *T* and performance measure *P*, if its performance at tasks in *T*, as measured by *P*, improves with experience *E*.[55]

The application of this definition in a military context can be illustrated with reference to an image classification task (*T*) such as determining whether a person is carrying a weapon. In such a case, the aim of the task is to increase classification accuracy (*P*) by training on labelled examples of each image class (*E*), such as people who are and people who are not carrying weapons.

There are many different ways in which machine learning programs can learn from experience or from training examples, but many of them involve updating some form

of weights. These weights are numbers that can be adjusted to change the influence of particular features of the input in creating the correct output. For example, if we are attempting to find a line of best fit $y = ax + b$, a is a weight that can be adjusted to change the influence of x on y so that our line fits our examples well. With standard machine learning, we often need to do very careful preliminary processing of the input data so that it better fits the learning program. For example, if we are learning $y = ax + b$, we need to ensure that we process x so that y increases a set amount as x increases, then our learning program can find that set amount a. If y increased with the square of x, we would need to take the square root of x before applying our learning program.

With traditional machine learning models, domain experts (i.e., those with specialised knowledge and expertise in the task the model is being applied to) hand-design rules to pre-process the data before the model learns the final prediction task based on that representation. This traditional method works well when the transformation is obvious; this is the case with our example above, where if y increases with the square of x, we just need to take the square root before applying the learning program. However, it is not an effective method for dealing with more complex cases—for example, determining from raw red, green and blue pixel values in images whether or not a person is carrying a weapon.

Neural networks are a type of machine learning model loosely based on the structure of biological brains. The difference between neural networks and other machine learning models is that, when trained well, neural networks can also learn the preliminary processing of the input data (pre-processing). Neural networks, as shown in Figure 1, have multiple layers of learned weights which allow them to take multiple steps both to pre-process the input data and to learn what the final prediction should be.

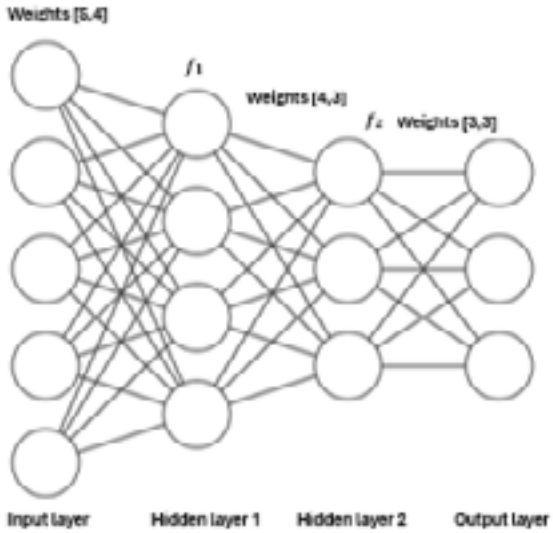


Figure 1: A neural network with three layers of weights

Deep learning is an application of neural networks with more than two layers of trainable weights. Technically a neural network of two layers of weights and unlimited weights in the first layer is a universal approximator, meaning it is able to predict any mathematical function as closely as we like if we keep adding more weights. However, in reality, this would not usually be possible to train, so neural networks with multiple hidden layers in task-specific architecture designs are used to encourage the model to learn. An example of this is a convolutional neural network, which is often used for predictions based on images.

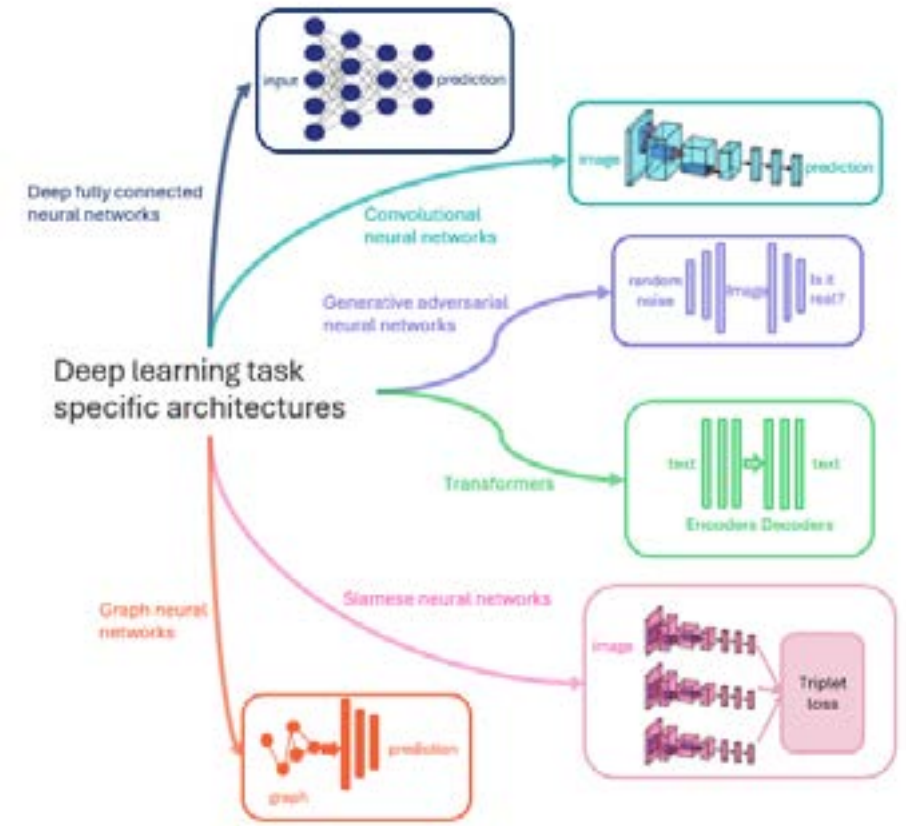


Figure 2: Different deep neural network architectures for different tasks

Modern computer vision models use neural networks (or deep learning) to make predictions from images, video and related data. In this way, complex relationships can be modelled between the input and desired output. In circumstances where the training data closely matches the testing data, top-performing deep learning models often produce better than human performance results for specific complex image-based prediction tasks. They are also significantly faster than humans at processing information and making predictions. However, some inherent challenges could potentially affect their use in military contexts:

1. Interpretability: Deep learning models have millions or even billions of parameters arranged in complex architectures, so it is difficult to pinpoint exactly what conditions cause a particular outcome.

2. Biases: Deep learning models are trained on data produced by humans, which is inevitably informed by human bias (such as racial discrimination), so they learn these biases. This bias has been evident in situations where computer vision has performed significantly worse on facial recognition for minorities. Joy Buolamwini and Timnit Gebru, in particular, highlight the compounding effects of intersectional discrimination (i.e., discrimination where protected characteristics like race and gender overlap) in widely used facial recognition algorithms.[56] More recent studies have similarly shown dramatic accuracy differentials in these kinds of applications between gender, age and racial groups—with non-dominant subpopulations suffering from the highest levels of misidentification, up to hundreds of times higher than dominant populations.[57]

3. Robustness: Deep learning models are only as good as the data on which they are trained. If there are major discrepancies between training data and real-world scenarios, performance can quickly degrade.

4. Security: Theoretical and practical weaknesses in deep learning models mean they are vulnerable to attack. Deep learning models learn in a bottom-up way, so they are vulnerable to low-level changes in data being used for prediction. For example, small, imperceptible perturbations in red, green and blue pixel values in images can cause major differences in predictions. These small changes can be deliberately created to mislead a model. For example, small stickers placed on stop signs can lead self-driving cars to misidentify the signs.[58]

There is growing awareness of AI vulnerabilities and flaws, not only among scholars and defence commentators but also within popular discourse regarding a broad range of AI applications.[59] Indeed, these limitations are part of the reason why the risks of over-reliance and moral deskilling are so serious. Even if these technologies were infallible and impenetrable, there would still be dangers associated with their use on the battlefield; the dangers associated with 'moral buffering', for example (see Section 3). But the dangers are all the more pronounced given that vision AI is itself prone to bias and error. It is for this reason that there are frameworks for the responsible use and development of (military) AI that (partially) address these issues. For example, the Method for Ethical AI in Defence developed by the Defence Science and Technology Group delivers an evidence-based ethical methodology for AI projects in Defence. It calls for AI that can be trusted and has clear lines of human responsibility for its development and use.[60]

/ 6. METHODOLOGY

6.1 Experiment Design

Both the ethical risks and the potential benefits of using vision AI on the battlefield rest on the same presupposition: that human decision-makers will, in fact, modify their behaviour in response to the information that AI presents or the labels these tools introduce into a visual field. However, this presupposition has not yet been empirically tested. Our experiments were designed to fill this gap. To this end, we created decision-making scenarios which we then overlayed with labels mimicking those that would ordinarily be produced by a vision AI decision support tool such as Athena AI, which combines AI computer vision, AI-enabled decision support, and information displays in a user interface.[61]

Our volunteer participants were military cadets / trainee officers from ADFA. These participants were recruited after we disseminated information about the study and the associated experiments within their respective ADFA divisions. We filmed a set of threat scenarios that these participants would, during the experiments, observe on a computer screen while controlling a crosshair with a mouse. The participants were instructed to decide whether and when to 'shoot' at the perceived originator of the threat simply by moving the crosshair (representing the sight of a rifle) and left-clicking. In some instances, the footage viewed was unadulterated (i.e., it contained no labels, thus inviting participants to use their own judgement regarding the threat). In other instances, text labels were overlaid on the screen mimicking those that would be generated by vision AI software. While some of these labels were correct (i.e., they contained accurate information), other labels were incorrect. In each scenario, there was presumed to be a 'correct' decision—i.e., a decision that conformed to the laws and customs of armed conflict as enshrined in Australia's rules of engagement. Our objective was to measure what influence, if any, the mock labels would have on participants' decision-making. Would they increase or decrease the likelihood of the 'correct' decision being made?

We created two series of video clips (Clip Series 1 and Clip Series 2). Each series included 12 different variations (scenarios), for a total of 24 scenarios. Role players acted in these scenarios. Clip Series 1 features a person of unknown affiliation or intention emerging from a cabin and approaching an insider—a dismounted member of a reconnaissance mission. The three variable features are outfit (civilian or military), item held (rifle or camera) and AI overlay (correct label, incorrect label or no label). These combinations produce 12 unique scenarios. Clip Series 2 follows the same general theme: a person of unknown affiliation or intention enters the frame and walks past an

insider. Here, the three variable features are gender (male or female), item held (phone or gun) and AI overlay (correct label, incorrect label or no label).

To shoot or not to shoot is essentially a choice. Our experiments were therefore a type of choice experiment, also known as a conjoint analysis.[62] This method was originally developed in the marketing discipline and has been applied in fields such as health care, economics, finance and technology. With the rise of AI-enabled decision support systems, we were particularly interested in the potential of AI to influence people's decisions. Therefore, our experiments aimed to establish the effects of factors that drive people's decisions. To achieve this, we systematically manipulated several independent variables in the series of video clips.

Our original design called for only one round of experiments. However, as outlined in Section 6.4, the results from the first round of experiments prompted us to make changes to the video clips and to conduct a second round. The first experiment round was conducted at an ADFA lab on 21 August 2024. The second round took place at the same location on 18 and 25 September 2024.

The Departments of Defence and Veterans' Affairs Human Research Ethics Committee approved the research, including all experimental procedures (approval # 474-22).[63] This ensured that our activities were carried out ethically and that we protected the mental and physical welfare, rights, dignity and safety of each volunteer. During the conduct of the experiments, participants were supported by training staff at all times.

6.2 Demographics

Fifty-four participants engaged in the first round of experiments. The second round comprised 27 participants. Both rounds of experiments were conducted independently and there was no overlap of participants. The volunteers, from all services, had completed at least their first year of study at ADFA; the majority were in their third year. Most of the participants identified as male, reflecting the gender distribution of the broader cohort from which they were recruited.

After each experiment round, we asked the participants to complete a questionnaire including information about their backgrounds (see Table 1). The sample compositions of the rounds were similar except for three statistically significant differences regarding the year of study at ADFA, completion of the Introduction to Military Ethics course (ZGEN2240) and experience with similar scenarios.

Table 1: Participants' backgrounds

		Round 1	Round 2
Total		54	27
In which year at ADFA are you?	2nd year	0	12
		0%	*44.4%*
	3rd year	52	15
		96.3%	55.6%
	4th year	2	0
		3.7%	*0%*
How do you describe your gender? gender?	Man/male	40	22
		74.1%	81.5%
	Woman/female	12	5
		22.2%	18.5%
	Non-binary	1	0
		1.9%	0%
	Prefer not to answer	1	0
		1.9%	*0%*
What service are you in?	Air Force	20	11
		37.0%	40.7%
	Army	17	10
		31.5%	37.0%
	Navy	17	6
		31.5%	*22.2%*
Were you familiar with the Red Card before today?	Yes	40	18
		74.1%	*66.7%*
	Somewhat	12	8
		22.2%	29.6%
	No	2	1
		3.7%	*3.7%*
I have completed ZGEN2022 Introduction to Military Ethics during my academic studies at UNSW Canberra.	Yes	31	9
		57.4%	33.3%
	No	23	18
		42.6%	*66.7%*
I would benefit from further learning on legal and ethical obligations in the conduct of war such as Rules of War Just War Tradition and International Humanitarian Law.	Yes	19	11
		35.2%	*40.7%*
	Somewhat	26	11
		48.1%	40.7%
	No	9	5
		16.7%	*18.5%*
I have experience with scenarios like the ones shown in the video clips.	Yes	6	9
		11.1%	33.3%
	Some	14	8
		25.9%	29.6%
	No	34	10
		63.0%	*37.0%*
I have experience with operating a remote weapons system (in training or in mission).	Yes	9	1
		16.7%	*3.7%*
	No	45	26
		83.3%	*96.3%*
I have combat experience.	Yes	3	1
		5.6%	3.7%
	No	51	26
		94.4%	*96.3%*
I have played computer games that have shooting settings.	Yes	44	24
		81.5%	88.9%
	No	10	3
		18.5%	11.1%

We used the same administrative procedures to support both rounds of the experiments. Members of the research team welcomed participants on arrival. Over a period of around 10 minutes, we then briefed them on AI-assisted target recognition in military combat environments (including limiting factors), presented background information concerning the experiments, and issued instructions (including a Q&A component). Following this, the experiment session itself was conducted, lasting approximately 20 minutes. This period included participants completing a post-experiment questionnaire.

6.3 Participant Briefing

During the briefing, the importance of positive identification of hostile intent was stressed to the participants. We also provided them with the following scenario and role information:

- You are a *gunner at a remote weapon system*, and you are protecting dismounted personnel who are on a reconnaissance mission. Your colleagues appear on the screen near the *tank in the foreground on the right*;
- Your role is to *provide cover fire for your colleagues*. You are seated in a protected shelter 500 metres distance from the checkpoint. You have a camera, screen, and a *remote control (mouse)* with fire buttons to control the gun;
- The area of the reconnaissance mission is known to be *hostile*, with civilians carrying weapons. This does not identify an enemy threat force. *Civilians and armed warriors* from opposing forces operate in the area simultaneously;
- Your order for opening fire is *a positive identification of hostile intent*. The *Red Card* is shown during the experiment.[64]

The participants were then given instructions for the experiment:

- There are two different *settings* in the clips; each setting has 6 different '*scenarios*' that are *independent* and (slightly) *different* in terms of:
    - The presence/absence of AI *labels* (which can be *correct* or *incorrect*);
    - The *words* in the AI labels;
    - The actions and/or appearance of the *other* person emerging in the clips.
- Your *decision* in each of the 12 clips:
    - You *shoot* the person emerging (and the clip finishes at that point); OR
    - You *do not shoot* the person (and the clip finishes by itself).

- After each of the 12 clips:
    - Please answer *two short questions* presented, then move to the next scenario.

Figure 3: Experiment 1 with correct AI label assigned to image

The two experiment settings in each round involved three different decision factors. In Experiment 1, those factors were the AI label (correct label shown; incorrect label shown; no label shown), the device carried by the person emerging (camera; rifle) and that person's clothing (military uniform and kit; civilian attire). In Experiment 2 (which included the same three AI labelling options), the additional factors were the person's device (phone; gun) and the person's gender (female; male). Given the number of factor categories involved, there were 12 different combinations for each experiment setting, with each combination represented in a different video clip (for a total of 24 clips). That allowed the identification of the individual effect on the shoot/no-shoot decision for all factor categories—e.g., the effect of incorrect labelling and the effect of the person's device.

Figure 4: Experiment 2 with correct AI label assigned to image

6.4 Assessment Considerations

Each participant was shown a total of 12 scenarios comprising six video clips from each of the two experiment settings. To militate against any learning or fatigue effects,[65] the scenarios were presented to the participants in a random fashion. The decision to adopt this process was based on an orthogonal design.[66]

After the first round of lab experiments concluded on 21 August 2024, we reviewed the initial results. They showed that AI labelling had no effect on shooting decisions. We surmised that the reason for this was that the actions of the role players and their weapons may have been too clearly visible on all occasions. In other words, the clarity of the video clips (clear vision) made the labelling less relevant as a potential decision factor. We decided to explore this hypothesis by reducing the clarity of the video clips in a second round of experiments. This was achieved by reducing the footage's brightness and resolution (blurring the footage). We assessed that this change would potentially add greater uncertainty around an item on the screen, and an AI label could potentially be a decision factor.

The last part of each video clip addressed the issue of 'hostile intent'. Specifically, the emerging person produced a relevant device (camera or rifle in the first experimental setting and phone or gun in the second). For the purpose of assessing the participants' responses, we identified that a 'correct' decision was when they 'shot' the person who emerged holding either a rifle or the gun and withheld shooting when a camera or phone was held. Conversely, a decision was incorrect when a participant shot the person who emerged holding a camera or phone and did not shoot when a rifle or gun was held.

Since the dependent variable 'correct shoot / incorrect shoot' is binary, we adopted a binary logistic regression model of statistical analysis to assess our results. This model is a type of 'repeated measures' analysis that can account for decisions made by the same participants under different conditions. The correct decision was modelled both as a linear function of the independent variables and as a constant. The technique of effects coding was used to code the categorical independent variables across the experiments (i.e., AI labelling, device, clothing and gender).

/ 7. FINDINGS

7.1 Data Analysis

The four models in Table 2 correspond to show the two experimental rounds (i.e., 'Clearer vision' and 'Poorer vision') and the two versions of the experiment (i.e., 'Clip Series 1' and 'Clip Series 2'). In each model, the dependent variable represents whether a correct decision was made. Table 2 summarises the modelling results. They are presented in the standard modelling format commonly used in discrete choice analyses. The values shown are parameter estimates for each of the variables, with standard errors reported in parentheses. Levels of statistical significance (p <.05, p <.01, p <.001) are indicated by asterisks.

Importantly, the results show that AI labelling was not a significant factor in any of the results. Instead, in Clip Series 1 (both clear and blurred images), participants' shoot/no-shoot decisions were significantly influenced by the device carried by the emergent figure. Further, clothing was a significant decision factor in the clear images in Clip Series 1. For instance, the highly significant parameter of –1.72 for the variable 'device' shows that when the device was switched from a rifle to a camera, the participant was less likely to make a correct decision. In Clip Series 2, the device was a significant decision factor in the poorer vision setting, while gender was significant in both settings.

Table 2: Model estimation results

Clip Series	Variable	Clearer vision	Poorer vision
Clip Series 1	Intercept	0.84 (0.17)***	–0.14 (0.21)
	Device (–1: Rifle, 1: Camera)	–1.72 (0.18)***	–1.67 (0.22)***
	Clothing (–1: Military clothing, 1: Civilian clothing)	–0.27* (0.15)	0.02 (0.21)
	AI labelling (–1: Incorrect, 1: Correct)	0.00 (0.21)	0.14 (0.3)
	AI labelling (–1: Incorrect, 1: No label)	0.14 (0.21)	–0.55 (0.31)
	R2	**0.36**	**0.37**
Clip Series 2	Intercept	2.98 (0.28)***	2.94 (0.44)***
	Device (–1: Gun, 1: Phone)	–0.29 (0.25)	–1.06 (0.39)**
	Gender (–1: Female, 1: Male)	0.55 (0.27)*	0.82 (0.34)**
	AI labelling (–1: Incorrect, 1: Correct)	0.44 (0.39)	–0.27 (0.4)
	AI labelling (–1: Incorrect, 1: No label)	–0.15 (0.34)	0.29 (0.43)
	R2	**0.05**	**0.17**

Note: ***, ** and * represent significance at .001, .01 and .05 levels respectively.

The parameter estimates of Table 2 are not directly comparable across the two experiment rounds; therefore, we adopted an elasticity-based methodology. Specifically, elasticity was measured as the change in the probability of making a correct decision in response to a change in one of the independent variables, divided by 2 (reflecting the difference in the effects coding values). To support this analysis, a base case was chosen for each clip series:

- Clip Series 1: The device is a rifle, the clothing is military clothing and the labels are incorrect. Under these conditions, the probabilities of making the correct decision for this base case are 93.69 per cent in the clearer vision experiment and 87.26 per cent in the poorer vision experiment.

- Clip Series 2: The device is a gun, the gender is male and the labels are incorrect. Under these conditions, the probabilities of making the correct decision for this base case are 91.85 per cent in the clearer vision experiment and 95.90 per cent in the poorer vision experiment.

The results of the analysis are shown in Table 3. The figures highlighted in bold correspond to the statistically significant coefficients in Table 2.

Table 3: Changes in probability of making correct decisions

		Clearer vision			Poorer vision		
		Probability base	**Probability variant**	**Elasticity**	**Probability base**	**Probability variant**	**Elasticity**
Clip Series 1	Switching from rifle to camera	93.69%	32.14%	**–30.77%**	87.26%	19.55%	**–33.86%**
	Switching from body armour to civilian outfit	93.69%	89.61%	**–2.04%**	87.26%	87.75%	0.24%
	Switching from incorrect label to correct label	93.69%	93.69%	0.00%	87.26%	89.97%	1.36%
	Switching from incorrect label to no label	93.69%	95.11%	0.71%	87.26%	69.60%	–8.83%
Clip Series 2	Switching from gun to phone	91.85%	86.37%	–2.74%	95.90%	73.77%	**–11.06%**
	Switching body from female to male	91.85%	97.14%	**2.64%**	95.90%	99.17%	**1.64%**
	Switching from incorrect label to correct label	91.85%	96.46%	2.30%	95.90%	93.21%	–1.35%
	Switching from incorrect label to no label	91.85%	89.37%	–1.24%	95.90%	97.67%	0.89%

In these results, the highest elasticity is observed for the ‘device’ variable in Clip Series 1. This result indicates that the ‘switching from a rifle to a camera’ scenario significantly reduced the likelihood of the participant making a correct decision. We consider that this outcome is likely to be due to the unusual shape of the camera. It is notable too that the elasticity for this variable is higher in the poorer vision experiment, suggesting that identifying the device becomes even more challenging under imperfect conditions. Switching from military to civilian clothing also reduced the likelihood of a correct decision being made, although only to a small extent. This result indicates that it may be slightly more difficult for remote weapon operators to make a distinction between subjects in military and civilian clothing. In Clip Series 2, switching from a gun to a phone

(with poorer vision) and switching from female to male decreased the likelihood of the participant making a correct decision. It should be noted, however, that most participants in this study identified as male, which may have influenced the observed gender-related effects. A more gender-balanced participant sample may have yielded different results. Exploring this issue further, however, is beyond the scope of the study.

During the experiments, we carried out several descriptive statistical analyses on participants' decisions. We also analysed their responses to post-scenario and post-experiment questions. The purpose of this exploratory approach was to provide deeper insights into how AI labelling might influence participants' decision-making.

Figure 5 illustrates the distribution of correct decisions across the clearer and poorer vision experiments. In both cases, the probability of making a correct decision was notably higher than the likelihood of making an incorrect decision. This outcome indicates that the scenarios were generally straightforward for participants. As anticipated, this probability decreased from 77.6 per cent in the clearer vision experiments to 69.4 per cent in the poorer vision experiments, indicating that participants found these scenarios to be more challenging. This result is notable because, in real life, military members are more likely to operate in poorer (rather than optimal) vision conditions.

Figure 5: Distribution of correct decision shown separately for each experiment

Figures 6 and 7 provide further detail about the distributions shown in Figure 5. Specifically, Figure 6 breaks down the distribution into shooting and withholding decisions. A key insight from this figure is that participants were more likely to make the correct decision when the correct action was to withhold fire rather than to shoot. In the experiments, the likelihood of making the correct decision when the correct decision was to withhold fire was 96.1 per cent for the clearer vision experiment and 86.0 per cent

for the poorer vision experiment. These probabilities reduced to 68.9 per cent and 62.8 per cent respectively when the correct decision was to shoot. This outcome suggests a tendency among participants to hesitate when it comes to pulling the trigger.

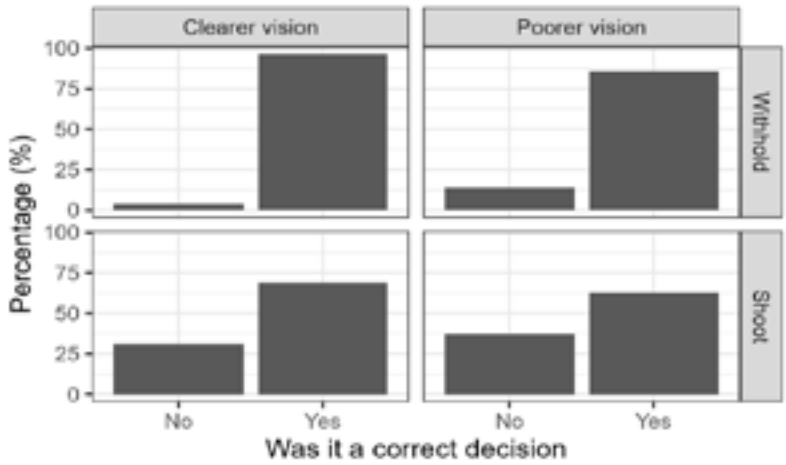


Figure 6: Distribution of correct decisions shown separately for each experiment and participant decision

Figure 7 presents the distribution of correct decisions for Clip Series 1 and 2 separately. It shows that the majority of incorrect decisions occurred in Clip Series 1 (38.9 per cent for the clearer vision experiment and 51.0 per cent for the poorer vision experiment).
Clip Series 1 was the scenario that featured the unusually designed camera. This object was chosen because, in the real world of military operations, objects may enter the scene that are novel to deal with both for human participants and for AI.

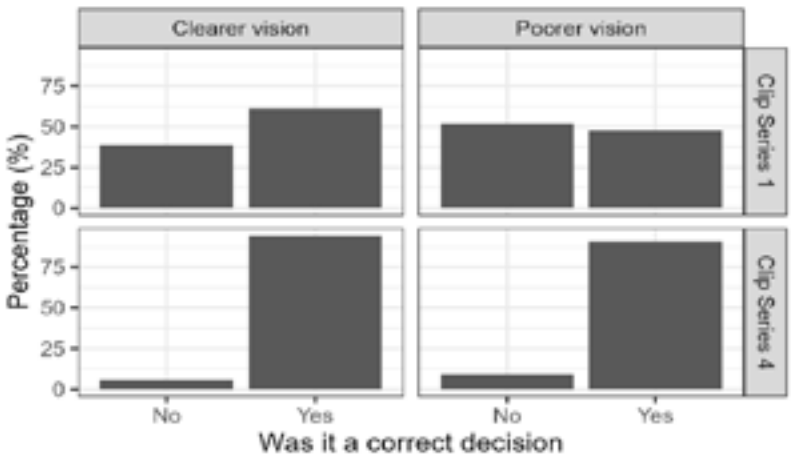


Figure 7: Distribution of correct decisions shown separately for each experiment and clip series

The distribution of correct decisions can be plotted separately for different AI labelling statuses. Specifically, Figure 7 shows that participant behaviour was noticeably different across Clip Series 1 and 2, so Figures 8 and 9 separately plot the AI labelling distributions. Figure 8 illustrates the results for Clip Series 1. In the clearer vision experiment, the highest percentage of correct decisions occurred when there was correct labelling (78.6 per cent), followed closely by no labelling (78.2 per cent) and then incorrect labelling (75.9 per cent). In the poorer vision conditions, however, the highest percentage of correct decisions was observed under incorrect AI labelling (72.2 per cent), followed by correct AI labelling (69.4 per cent) and then no labelling (66.7 per cent). Figure 9 presents the corresponding plots for Clip Series 2. Here, in most cases, the participants made the correct decision, and there were negligible variations in outcomes across different labelling conditions.

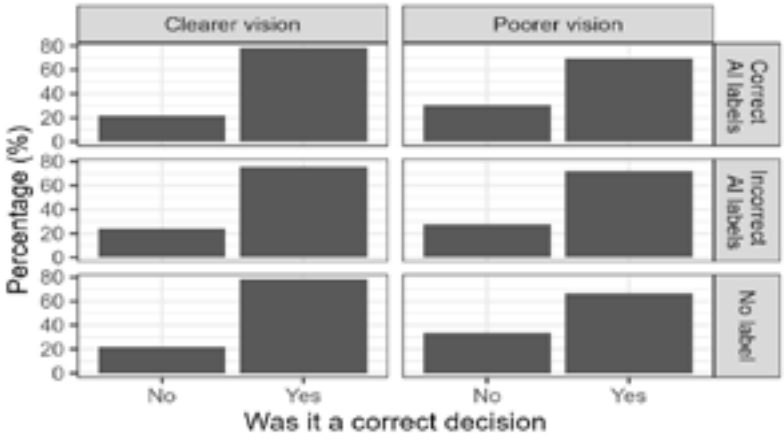


Figure 8: Distribution of correct decisions for Clip Series 1 shown separately for each experiment and AI labelling

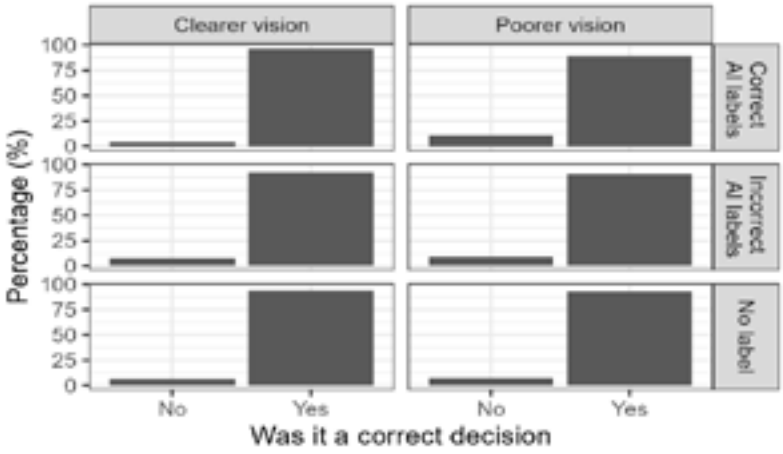


Figure 9: Distribution of correct decisions for Clip Series 2 shown separately for each experiment and AI labelling

After making their decision in each scenario, participants were asked to respond to the following two questions:

1. *I feel I made the right decision to either shoot or withhold fire*.
   Response options: Disagree / Neither agree nor disagree / Agree
2. *How do you feel the labels on the screen affected your decision-making?*
   Response options: Not applicable / Positively / Negatively / Not at all.

Figure 10 illustrates how participants responded to the first question, with the data divided by correct or incorrect decisions and the two rounds of experiments (clearer and poorer vision conditions). While confidence levels (i.e., how confident the participants were that they had made the correct decision) were generally higher when participants made correct decisions, they remained overwhelmingly high even when participants' decisions were incorrect. Interestingly, confidence levels were notably higher in response to the poorer vision experiment compared to the clearer vision experiment.

Figure 11 presents the responses to the same confidence question, this time categorised by AI labelling status and the two experimental rounds. This figure further underscores that participants maintained high levels of confidence across the three levels of AI labelling. This outcome is consistent with the modelling results, which showed that the three labelling options did not affect the likelihood of making the correct decision. It also suggests that participants trusted their own interpretations of what they saw in the video clips. Remarkably, even in the poorer vision experiment, and when AI labels were incorrect, participants remained confident that they had made the right decision.

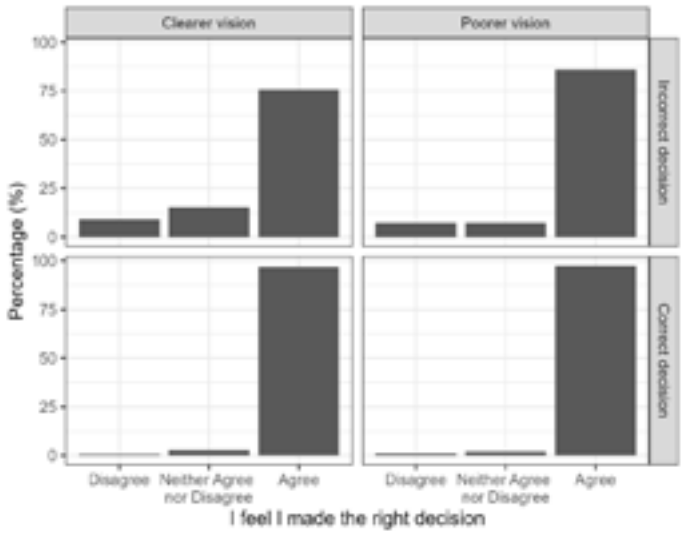


Figure 10: Distribution of participants' perceived correctness shown separately for actual correctness and each experiment

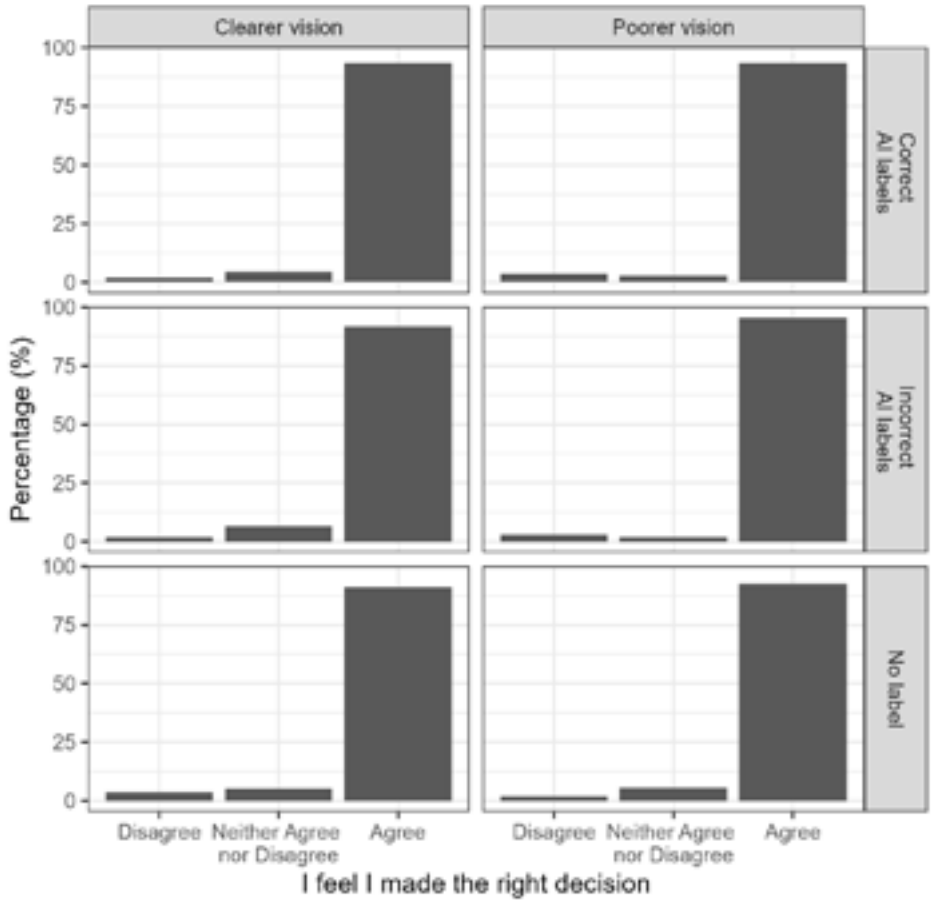


Figure 11: Distribution of participants' perceived correctness shown separately for AI labelling and each experiment

Figure 12 illustrates the distribution of responses to the second post-scenario question: 'How do you think labels affected your decision?' The figure is segmented by AI labelling status and by experiments. In the clearer vision experiment, and with correct AI labels, 51.8 per cent of participants reported that the labels had no effect on their decisions, followed by 29.1 per cent who noted a positive effect. Meanwhile, 19.1 per cent experienced a negative effect. This outcome suggests that when visibility is good and it is relatively easy to distinguish threats from non-threats, correct AI labels either are mostly ignored or serve to confirm what participants already believe. The fact that 19.1 per cent of participants reported that the AI labelling negatively affected their decision-making is likely to stem from their having interpreted the situation incorrectly. This misperception led participants to perceive the AI labels as contradicting their understanding.

For incorrect AI labels in the clearer vision experiment, participants most commonly reported that this factor did not affect their decision-making. Negative effects rose to 34.2 per cent, probably reflecting participants who correctly assessed the situation but were confused by the contradictory AI label. The fact that 19.4 per cent of participants reported that incorrect AI labels had a positive effect on their decision-making is likely to stem from their having incorrectly analysed the situation, with the incorrect label reinforcing their mistaken conclusions.

In the poorer vision experiment, the 'no effect' category declined significantly. For correct AI labels, the most frequently reported effect was positive (37.7 per cent), while for incorrect AI labels, it was negative (45.2 per cent). These findings indicate that when visibility is impaired and object identification becomes more challenging, participants are more reliant on AI labels. Another noteworthy observation is the increased proportion of negative effects for correct labels (24.5 per cent) and positive effects for incorrect labels (22.1 per cent). These cases probably arose when participants incorrectly assessed the situation, leading to contradictions between their evaluation and the correct AI label (or alignment of their decisions with an incorrect label).

This analysis highlights that participants tend to prioritise their own judgement over AI labels. Labels that confirm participants' conclusions, whether correct or incorrect, are more likely to be perceived positively. Conversely, labels that contradict their judgements are more likely to be perceived negatively, even if the labels are accurate.

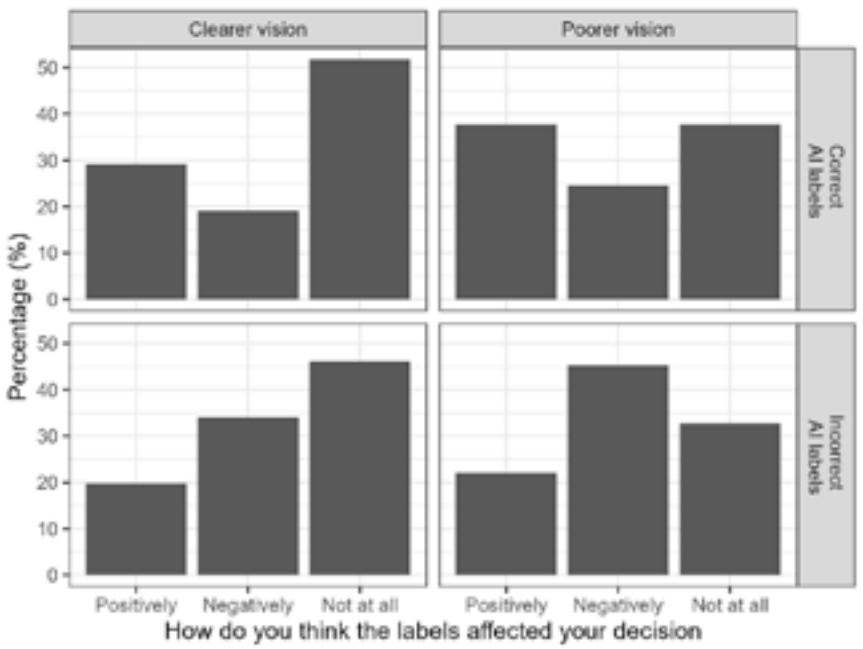


Figure 12: Distribution of AI labelling effect shown separately for AI labelling and each experiment

The post-experiment questionnaire asked participants to assess the realism of the scenarios: *Overall, do you think the scenarios were represented realistically?* Response options were Yes, Mostly, Somewhat and No.

Figure 13 illustrates the distribution of participants' responses to this question. The results show that in both cases, over 50 per cent of participants perceived the scenarios as either completely or mostly realistic. Notably, this percentage was even higher in the poorer vision experiment, suggesting that reduced visibility may have enhanced the perceived authenticity of the scenarios. These results suggest that the experimental approach served its purpose.

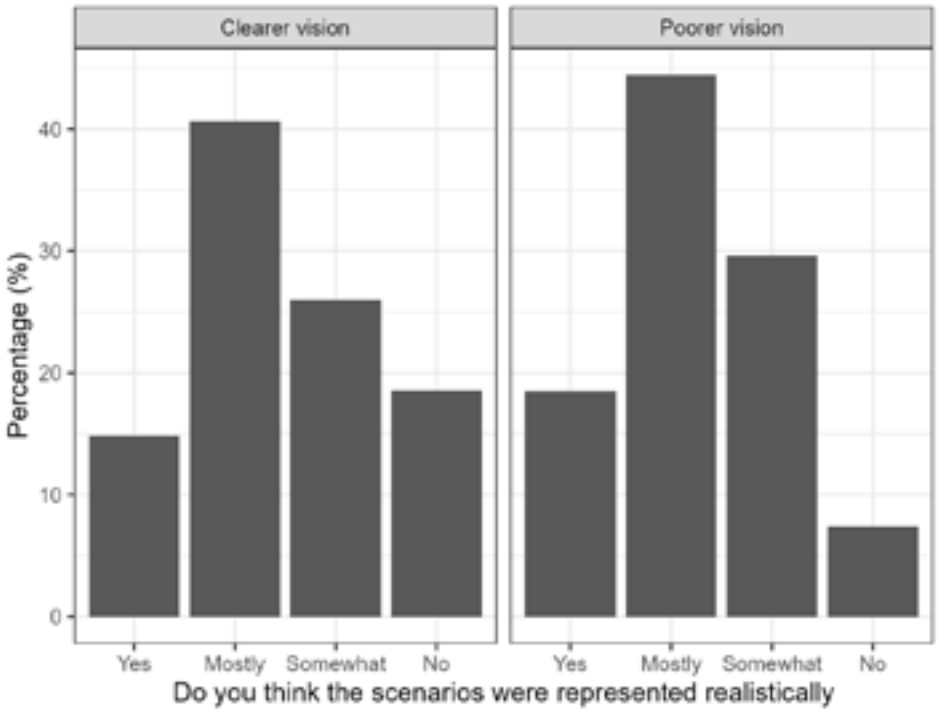


Figure 13: Distribution of participants' responses to the question regarding realistic portrayal

Feedback from the participants was collected from post-experiment questionnaires. These questions were open-ended, and the participants were able to provide detailed comments on their experience and perception of the scenarios and AI labels. We employed thematic coding, a qualitative data analysis technique, to analyse this feedback.

7.2 Participant Observations

For our initial coding, we identified recurring themes from the participant feedback, such as realism, trust in AI, decision-making dynamics and ethical considerations. A number of sub-themes, such as realism of scenarios and perceived lack of urgency, were also apparent. We assigned codes to specific phrases or sentences representing these themes and sub-themes. For our thematic analysis, we grouped the codes into broader categories, such as 'experimental scenario design', 'trust and effect of AI labelling' and 'ethical and operational concerns'.

This section considers the most significant themes that arose from participant feedback on the experiments. These themes were realism, temporal dynamics, trust in AI, and human oversight.

Realism

As indicated by Figure 13, most participants found the individual scenarios to be sufficiently realistic. However, the repeated nature of the experimental design did contribute to a broader sense of unrealism. Specifically, the repetitive use of certain actors and scenarios, along with recycled video clips (with minor changes in correct and incorrect labelling, clothing and gender), undermined the authenticity and immersive quality of the experiments. As we have previously noted, the enemy was always obvious and behaved in a distinctly hostile manner. This felt unrealistic to some participants, who considered that a real opponent would display more stealthy or unpredictable behaviour. Friendlies, enemies, weapons or objects of interest were also clearly visible or apparent from the beginning, as were AI labels (when used). In addition, there was considerable visual similarity between scenarios; some respondents remarked that they could not readily distinguish between them.

Temporal Dynamics

Participants consistently commented on the pace of the experiments (each video clip taking place over a 20-second timeframe). A number of participants felt it was excessively prolonged. For those who felt it was drawn-out, there was no sense of urgency. This aspect of the experiment design warrants further research, as the pacing of the scenarios may have influenced performance. Specifically, because there was ample time for situational assessment, less reliance may have been placed on AI labelling than if participants had perceived more urgency in their decision-making.

Trust in AI

As part of the experiment design, we deliberately introduced erroneous AI classifications into the scenarios. Many participants reported that when they detected the first mistake, they did not trust the AI in subsequent scenarios. This loss of confidence in AI highlights the fragility of trust in AI and autonomous systems.

Human Oversight

Many of the participants mentioned ethical considerations. They emphasised the importance of retaining human oversight in critical decision-making processes. Perhaps linked to the loss of trust in the AI, participants stated that the acknowledged and perceived limitations of AI made them focus on factors an automated system would struggle to interpret, particularly non-verbal cues such as body language and movement.

7.3 Limiting Factors

The participant feedback themes of realism, temporal dynamics, trust in AI, and human oversight all potentially influenced the experiment results. Analysis further identified other limitations arising from the fact that the experiments were based on lab simulations that did not reflect real life.

Use of Video Clips

The experiments' representations of persons and equipment were an attempt to maintain participant engagement and effectively simulate real operational scenarios, but participants identified shortcomings in the scenarios due to their lack of realism. To enhance perceptions of realism, future scenarios should feature more diverse role players. There should also be no repetition or reuse of role players throughout the scenarios. Given the demonstrated need for clear and poor vision scenarios, future scenarios should take into account environmental complexities, such as multiple moving entities, and should be created with reduced visual clarity. These changes would more accurately reflect the ambiguity and fluidity of real-life battlefield conditions, offering a more credible platform for studying human–AI interactions.

Despite these limitations, the inherent simplicity of our experiment design had some benefits. Specifically, it ensured that no confounding factors were introduced into the experiment scenarios. It is probably for this reason that both participant groups achieved similar results.

Length of Video Clips

The brevity of the video clips—20 seconds or so—ensured that participants were not cognitively tired throughout, or even at the end of the experiment. In real life, however, battlefield scenarios can last for hours. Fatigue in those circumstances is a very real factor in impaired decision-making. As noted earlier, the timing of the scenarios, the sense of urgency it may create and the potential effects of longer-term fatigue warrant further research.

Attire of Role Players

The attire of role players may have added to the lack of realism in the scenarios, thus affecting decision-making. As the insiders were not wearing uniforms that represented those worn by members of the ADF or allied forces, participants were unable to judge on which side the role players were acting.

Mistrust of AI Labels

As part of the experiment design, participants were shown random mistakes made by the AI classifier. While this reflected 'real AI' (which is not perfect in its predictions), participants reported distrust in subsequent classifications, even when the AI label was 'correct'. This observation reinforces the need to design experiments which incorporate more realistic and controlled AI reliability and failure rates. Setting AI's reliability at 80 per cent and introducing randomised errors could better simulate the real-world performance of AI systems and their imperfections. Furthermore, scenarios designed to show the confidence values of AI (to visually evaluate its errors) could help us better understand how trust in AI systems works.

Preconceptions

Vision AI is currently in use in the conflicts in Ukraine and Gaza. Both conflicts feature prominently in social and print media, TV reports and online articles. Therefore, participants' perceptions of the reliability (or otherwise) of vision AI may already have been shaped by their perceptions as to whether it has—or has not—contributed to just outcomes in these conflicts. More generally, with more vigorous reporting of the fallibility of popular AI tools such as ChatGPT and GROK, participants may have been culturally primed to mistrust any form of AI.[67]

8. DISCUSSION AND RECOMMENDATIONS

8.1 Discussion

Our results indicate that AI labelling did not have statistical significance in the models across both experimental conditions. Variables such as device type, clothing and gender (see Table 2, Clip Series 2) significantly influenced decision appropriateness, particularly under poorer vision conditions. For instance, in Clip Series 1, the transition from a rifle to a camera decreased decision accuracy, probably due to the unconventional shape of the camera. Similarly, the hesitancy observed in participants' shooting decisions (accentuated by lower decision accuracy in shooting scenarios compared to withholding fire) suggests a cautious approach driven by uncertainty, for which AI detection systems must account. The AI labels, by contrast, had no statistically significant effect. Although AI labelling showed no statistically significant effect on the likelihood of a correct decision being made, the experiments nevertheless revealed interesting underlying issues concerning the degree to which participants relied on AI labels. In this regard, the following points are particularly relevant:

- Quality of vision: In clearer vision scenarios, AI labels often acted as confirmatory tools (with most participants reporting no substantial impact on their decision-making), whereas under conditions of poorer visibility, participants' reliance on AI labels increased somewhat.
- Personal judgement: Participants exhibited higher confidence in their decisions even when the labels were incorrect. Interestingly, this confidence persisted across varying labelling conditions, underscoring a paradox whereby participants trusted their subjective assessment over contradictory AI inputs.
- Mistrust of AI: Once a participant had seen the AI misclassify a relevant item, the participant would mistrust the AI system from that moment onwards, even though it showed correct labels on subsequent occasions.

These findings highlight the need to improve AI–human interaction design. They suggest, too, that human users should be trained to balance their own judgement with their trust in AI systems. As with the introduction of any new technology, the art of habituation is a key factor for responsible use of the technology. The art of habituation includes an awareness of the limitations and potential dangers of using AI in certain safety-critical situations.

It is significant that participants in the experiments preferred to rely on their own individual interpretation of the situation rather than on an AI classification. This highlights the importance of military training and the underlying decision-making processes that enable individuals to interpret cues that are hard for AI to capture or classify. Such cues may include body language or facial expressions. This finding reinforces the importance of ensuring that autonomous weapons systems are always supported by 'human-in-the-loop' frameworks.

Real-world conditions and scenarios are inherently unpredictable, requiring dynamic, nuanced and rapid decisions. If AI is to be effectively integrated into remote weapons systems used in such environments, efforts will be needed to enhance reliability, transparency and contextual adaptability. These measures will help ensure that any AI support that is provided consistently aligns with human users'/operators' decision-making processes, particularly in high-stakes environments.

Further experiments could benefit from incorporating ethically ambiguous scenarios or hidden nefarious behaviour, forcing participants to weigh AI recommendations against their own judgement. This approach would allow for a greater understanding of the ethical boundaries of weapon autonomy and its impact on operators' behaviour.

8.2 Recommendations

Although we found that vision AI use has no significant positive or negative influence on choices, a number of limiting factors may have contributed to this finding. These include aspects of the experimental design and the various other limitations discussed in Section 7. Given these considerations, the emerging feedback themes, underlying patterns revealed in our modelling analysis, and potential additional influencing factors such as gender, we believe further research is needed to conclusively determine the effects of vision AI on ethical decision-making. We therefore recommend that further empirical studies be undertaken to determine whether AI-enabled support tools are likely to have any effect (positive or negative) in a wider variety of military contexts.

For future experiments we recommend:

- Extend the overall length of the experiment in terms of the scenarios (i.e., the total length of individual clips stitched together) presented to participants to provide sustained cognitive engagement. This could better mimic the continuous high-pressure environment faced by operators on the battlefield. Reducing clarity and visibility and increasing overall complexity with more objects of interest, friendlies and enemies may well encourage more explicit reliance on AI labelling.
  This approach would cater to both immediate reaction and longer-term strategic decision-making studies in the presence of AI labelling.

- Provide more detailed briefing to participants regarding the stakes involved in the decisions they are being asked to make. If the stakes are high, where incorrect decisions will have potentially fatal or deeply detrimental consequences, we can better assess if participants rely more heavily on AI for assistance or use it to diffuse their sense of moral responsibility—i.e., if they are more likely to shift 'blame' onto their technological aides under such conditions. In cases of over-trust/under-trust leading to mistakes, we could perhaps better assess the effects of moral emotions such as guilt and shame. Indeed, in future the design of such experiments should proceed on the assumption that AI systems are not neutral tools, but may be considered themselves 'moral environments' which affect the human operators' behaviour and reasoning, so that both positive and moral ethical behaviour can be reinforced.[68]
- Ensure the scenarios depict more 'realistic' battlefield conditions, including multiple moving entities, reduced visibility, and time-constrained decision-making, to better understand AI's impact under stress. The scenarios we presented lacked the ambiguity and complexity of real battlefield conditions, potentially reducing the reliance on AI labels. Furthermore, the experiments did not fully explore the dynamics of trust in AI systems over time, particularly how initial errors influence long-term operator behaviour.
- Ensure the design incorporates ethically ambiguous, suspicious, or more subtle and discreet nefarious behaviours, which would force participants to weigh AI recommendations against their own intuitive judgements. This approach would allow for a greater understanding of the ethical boundaries of weapon autonomy and its impact on operators' behaviour.
- Enhance training programs by incorporating scenarios with AI-assisted decision-making into the ADF curriculum in order to familiarise future and current operators with the shortcomings and limitations of such systems. As the nature of weapons systems changes, high-order critical thinking and situational awareness skills will be needed to mitigate the accompanying risks. Future training packages, for example, could include the potential for technological mediation of moral decision-making on the battlefield.
- Generate clear protocols that mandate human oversight, ensure accountability and reinforce adherence to the relevant ethical and legal standards. Such protocols must inform the design of weapons systems that have implications for ethical decision-making and moral deskilling. These standards should be developed through engagement with experts in military ethics, IHL and behavioural psychology.

/ 9. CONCLUSION

Several Australian Government agencies have recently published documents offering guidance on the development and use of AI. Among these is the Policy for the Responsible Use of AI in Government, published by the Digital Transformation Agency in September 2024. This policy underpins safe, ethical and responsible use of AI technologies by providing baseline requirements on governance, assurance and transparency. Another initiative is the Department of Industry, Science and Resources Voluntary AI Safety Standard. Accompanying this is the National AI Centre's AI Impact Navigator,[69] which facilitates safe and responsible adoption of AI and aims to help companies better understand the potential impact of AI systems on stakeholders. The generation of such guidance demonstrates the serious consideration currently being given to the practical application of emerging AI technology.

Of most relevance to the Australian Army is the Robotic & Autonomous Systems Strategy (RASS). This strategy outlines how Army intends to implement emerging interconnected technologies to maximise its combat advantage through improved training, fighting, and decision-making strategies. To achieve this aim, the RASS particularly focuses on AI, autonomy and robotics, including uncrewed systems, self-learning machines, and systems capable of making sense of their environment.[70] While the RASS underscores Army's commitment to the integration of AI into its capabilities, it currently lacks detailed focus and evidence concerning the ethical implications of deploying AI in critical environments. Notably, consideration of ethics is not prominent in the RASS.

Our research into and findings on AI and ethical decision-making provide much-needed empirical insights into integrating vision-based AI systems into the Australian Army's operational framework. Importantly, our findings address a gap in understanding of how AI-enabled decision support systems could influence ethical decision-making on a battlefield, particularly in a shoot/no-shoot scenario.

Whereas our pilot study suggested that AI labelling does have some influence on moral decision-making, this initial assessment was not supported by our experiments. Instead, the experiments suggested that vision AI has no statistically significant positive or negative influence on choices. While we have identified limitations in the experiment design, our findings nevertheless suggest that some of the ethical disquiet surrounding the introduction of vision AI to battle is not warranted, or is at least premature. Our analysis also suggests that some of the optimism about the potential for vision AI to improve ethical decision-making in battle may also be unfounded. Rather than having the negative effects feared or the positive effects hoped for, AI-enabled support tools in battle may have no effect to speak of, at least in the kinds of circumstances simulated in our experiments.

Our experiments have highlighted potential limitations in the effective use of vision AI. Gaining this knowledge about when and how technological mediation (as well as other factors such as social media and personal world views) influence soldiers' decisions is important. Not only does it help ensure ethical conduct on the battlefield; it also supports the responsible operational use of AI in current and future conflicts. Our research is therefore relevant for military ethics education.

Much of the debate regarding the ethics of AI in military operations focuses on the application of lethal autonomous weapons systems. AI use in the military context, however, has a much broader scope (vision-based AI is an example of this), and there are considerable benefits to its correct application.[71] As HW Meerveld and colleagues urge, the opportunities available should not be ignored. Indeed, it would be irresponsible not to take advantage of such technology.[72] Our research, however, highlights the need for robust training of future soldiers to ensure that these opportunities are not forgone. While the research we have conducted to date is limited in scope, this work and our recommendations for continued inquiry make a direct contribution to the Army's ongoing efforts to modernise its capabilities while maintaining compliance with ethical and legal frameworks. It is through work like this that Army's efforts to integrate AI technologies will enhance, rather than undermine, ethical decision-making on the battlefield.

Insights from research such as ours, which focuses on the effects of technological factors on ethical decision-making, are the bedrock upon which ADF initiatives can be instigated, discussions fostered and guidelines developed. In this way, we hope that our research contributes to better ethical use of decision support tools that ensure minimal harm to the human user/operator while maintaining the ethical standards demanded of the ADF by government and by the Australian community at large.

/ ACKNOWLEDGEMENTS

We acknowledge that the Red Cross is a protected symbol, and a flag bearing this symbol may only be formally handed out by the ICRC. We did not have access to a formal Red Cross flag, and hence we created our own version for the sole purpose of experimentation with military AI software.

We wish to thank all ADFA personnel and trainee officers involved in this project as participants, role players and supporting personnel.

Funding

Funding for this research was granted through the Australian Army Research Scheme 2024.

/ ABOUT THE AUTHORS

Dr Christine Boshuijzen-van Burken was a researcher at the School of Systems and Computing at UNSW Canberra. She is currently at the Netherlands Defence Academy and Eindhoven University of Technology in the Netherlands. Her research interests include the ethics of (military) technology, military ethics, and reformational philosophy.

Dr Jo Plested is an associate lecturer at the School of Systems and Computing at UNSW Canberra who has been researching deep learning for 10 years. Her expertise is focused on transfer learning for small, specialised datasets. She is an expert in AI.

Dr Ned Dobos is associate professor in International and Political Studies at the School of Humanities and Social Sciences at UNSW Canberra. His research interests include military ethics, political philosophy, humanitarian intervention and moral injury.

Dr Twan Huybers is an associate professor at the School of Business at UNSW Canberra. His research interests are behavioural economics, choice experiments and choice modelling approaches.

Dr Milad Ghasrikhouzani is a senior lecturer at the School of Engineering and Technology at UNSW Canberra. His research interests include travel behaviour and transport modelling, discrete choice analysis, and consumer preference research related to transport innovation adoption.

Dr Erandi Lakshika Hene Kankanamge is a senior lecturer at the School of Systems and Computing at UNSW Canberra. Her research interests include multi-agent systems, computational intelligence, multi-objective optimisation, human–computer interfaces, and games for health.

Professor Deane Baker is professor of ethics in the School of Humanities and Social Sciences at UNSW Canberra. His research interests include military ethics, the ethics of lethal autonomous systems, and the privatisation of military forces.

Dr Oleksandra Molloy is a senior lecturer at the School of Science at UNSW Canberra. Her research interests include uncrewed and autonomous systems, human factors and aviation safety.

Abhi Veda is a higher degree research student at the School of Engineering and Technology at UNSW Canberra. His research interests include AI and robotic vision.

/ ENDNOTES

1 'AI Decision Making: What Is It, Benefits & Examples', *Intellias*, 3 January 2025, at: https://intellias.com/ai-decision-making; Oleksandra Molloy, *Drones in Modern Warfare: Lessons Learnt from the War in Ukraine,* Australian Army Occasional Paper No. 29 (Australian Army Research Centre, 2024), p. 1, at: https://doi.org/10.61451/267513.

2 'New Report on Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts', *Geneva Academy*, 13 May 2024, at: https://www.geneva-academy.ch/news/detail/716-new-report-on-artificial-intelligence-and-related-technologies-in-military-decision-making-on-the-use-of-force-in-armed-conflicts. This report particularly focuses on how AI technology affects decision-making and compliance with international humanitarian law. It also highlights concerns about the risk of AI-based systems, accountability and transparency.

3 Sarah Grand-Clément, *Artificial Intelligence Beyond Weapons: Application and Impact of AI in the Military Domain* (UNIDIR, 2023), at: https://unidir.org/wp-content/uploads/2023/10/UNIDIR_AI_Beyond_Weapons_Application_Impact_AI_in_the_Military_Domain.pdf; International Committee of the Red Cross Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects (Geneva, June 2005), at: https://www.icrc.org/sites/default/files/external/doc/en/assets/files/other/icrc_002_0811.pdf.

4 *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflict* (International Committee of the Red Cross, 2024), at: https://shop.icrc.org/international-humanitarian-law-and-the-challenges-of-contemporary-armed-conflicts-building-a-culture-of-compliance-for-ihl-to-protect-humanity-in-today-s-and-future-conflicts-pdf-en.html. Section 3 addresses AI in decision-making.

5 Peter Layton, 'Evolution Not Revolution: Defence AI in Australia', in Heiko Borchert, Torben Schütz and Joseph Verbovszky (eds), *The Very Long Game: Contributions to Security and Defence Studies* (Springer, 2024), pp. 581–603, at: https://doi.org/10.1007/978-3-031-58649-1_26.

6 *The Geneva Conventions of 12 August 1949* (International Committee of the Red Cross), at: https://www.icrc.org/sites/default/files/external/doc/en/assets/files/publications/icrc-002-0173.pdf.

7 Mariarosaria Taddeo and David Sutcliffe, 'The Ethics of Artificial Intelligence in Defence', *Oxford Internet Institute* (website), 3 February 2025, at: https://www.oii.ox.ac.uk/news-events/the-ethics-of-artificial-intelligence-in-defence.

8 Bernard Marr, 'How AI Is Used in War Today', *Forbes*, 17 September 2024, at: https://www.forbes.com/sites/bernardmarr/2024/09/17/how-ai-is-used-in-war-today.

9 Christine Boshuijzen-van Burken, Deane-Peter Baker, Ned Dobos, Milad Ghasrikhouzani, Erandi Hene Kankanamge, Twan Huybers, Oleksandra Molloy, Jo Plested and Shreyansh Singh, 'Understanding Ethical Implications of AI Enabled Decision Support Systems on the Battlefield', *Research Symposium (RSY): Meaningful Human Control in Information Warfare* (Amsterdam: NATO, 2025), at: https://www.sto.nato.int/document/understanding-ethical-implications-of-ai-enabled-decision-support-systems-on-the-battlefield.

10 Athena AI, at: athenadefence.ai.

11 Lieutenant Colonel James E Downey, U.S. Army War College, explored the apparently contradictory notion of military ethics, along with the scepticism with which some view it, in his study project *Professional Military Ethics: Another Oxymoron*? (Defense Technical Information Centre, 1989), at: https://apps.dtic.mil/sti/citations/ADA209233.

12 Valerie Morkevičius, *Realist Ethics: Just War Traditions as Power Politics* (Cambridge University Press, 2018), p. 261.

13 Deane-Peter Baker, *Key Concepts in Military Ethics* (NewSouth Publishing, 2015), passim.

14 The United National Charter of 1945. See for an outline 'What Are Jus ad Bellum and Jus in Bello?', *International Committee of the Red Cross* (website), at: https://www.icrc.org/en/document/what-are-jus-ad-bellum-and-jus-bello-0.

15 See for example Brian Orend, *War and International Justice: A Kantian Perspective* (Wilfrid Laurier University Press, 2000), note 8, p. 258; and David Michael Jackson, 'Jus ad Bellum', in Deen K Chatterjee (ed.), *Encyclopedia of Global Justice* (Springer, 2011), pp. 581–582.

16 It is generally recognised that the US invasion of Iraq in 2003 did not have a just cause basis; it was driven by motives other than self-defence. See for example Ting Chun Ngai, 'Was Iraq War a "Just War" or Just a War? An Analysis from the Perspectives of Just War Theory', *Open Journal of Political Science* 9, no. 2 (2019): 373–382, at: https://www.scirp.org/journal/paperinformation?paperid=91783; David DeCosse, 'Totaling Up; It Was an Unjust War', *Markkula Center for Applied Ethics at Santa Clara University* (website), 16 May 2008, at: https://www.scu.edu/ethics/focus-areas/more-focus-areas/resources/totaling-up-it-was-an-unjust-war.

17 Matthew Talbert and Jessica Wolfendale, *War Crimes: Causes, Excuses and Blame* (Oxford University Press, 2019).

18 'Afghanistan Inquiry', *Department of Defence* (website), at: https://www.defence.gov.au/about/reviews-inquiries/afghanistan-inquiry. The redacted version of *Inspector-General of the Australian Defence Force Afghanistan Inquiry Report* (Brereton Report) is available on this site.

19 Chief of the Defence Force, 'Press Conference—IGADF Afghanistan Inquiry', 19 November 2020, transcript at: https://www.defence.gov.au/news-events/releases/2020-11-19/press-conference-igadf-afghanistan-inquiry.

20 Brereton Report, 'Chapter 1.08: War Crimes in Australian History', pp. 183–241.

21 Rory Callinan, 'The Alleged Afghanistan War Crimes That Shocked Australia', *The Guardian*, 19 November 2020, at: https://www.theguardian.com/australia-news/2020/nov/19/the-alleged-afghanistan-war-crimes-that-shocked-australia.

22 Martin Crotty and Carolyn Holbrook, 'The Anzac Legend Has Blinded Australia to Its War Atrocities. It's Time for a Reckoning', *The Conversation*, 7 December 2020, at: https://theconversation.com/the-anzac-legend-has-blinded-australia-to-its-war-atrocities-its-time-for-a-reckoning-151022.

23 The Hon Richard Marles MP, 'Statement on the Closure of the Afghanistan Inquiry Report', 12 September 2024, transcript at: https://www.minister.defence.gov.au/statements/2024-09-12/statement-closure-afghanistan-inquiry-report.

24 The centre's remit is to provide advice, education and research to enhance command, leadership and ethics within the Department of Defence. It plays a substantial role in guiding Joint Professional Military Education throughout Defence. See 'Biography: CDLE', *The Forge*, at: https://theforge.defence.gov.au/author/cdle.

25 'ADF Philosophical Doctrine—Military Ethics', *The Forge*, at: https://theforge.defence.gov.au/military-ethics/adf-philosophical-doctrine-military-ethics.

26 Deane-Peter Baker, Rufus Black, Roger Herbert and Iain King, *Ethics at War: How Should Military Personnel Make Ethical Decisions?* (Routledge, 2024).

27 Peter-Deane Baker, an ethicist, is a philosopher by training. He is also a former serving officer in both the British and South African armies. Ned Dobos is a moral philosopher, and Christine Boshuijzen-van Burken holds Bachelor of Science degrees in human movement and mechanical engineering and a PhD in ethics and the philosophy of technology.

28 Baker, *Key Concepts in Military Ethics*.

29 Amanda Sharkey, 'Autonomous Weapons Systems, Killer Robots and Human Dignity', *Ethics and Information Technology* 21 (2019): 75–87, at: https://link.springer.com/article/10.1007/s10676-018-9494-0.

30 Samantha Crompvoets, *Special Operations Command Culture and Interactions: Perceptions, Reputation and Risk* (2016), at: https://www.defence.gov.au/about/reviews-inquiries/afghanistan-inquiry/resources. This report precipitated the Brereton Inquiry.

31 Stuart Jeffries, '*War and Justice: The Case of Marine A* Review—Compelling Account of a Killing, with a Fatal Flaw', *The Guardian*, 1 August 2022, at: https://www.theguardian.com/tv-and-radio/2022/jul/31/war-and-justice-the-case-of-marine-a-review-compelling-account-of-a-killing-with-a-fatal-flaw.

32 Max Bazerman and Francesca Gino, 'Behavioral Ethics: Toward a Deeper Understanding of Moral Judgment and Dishonesty', *Annual Review of Law and Social Science* 8 (2012): 85–104 (90), at: https://doi.org/10.1146/annurev-lawsocsci-102811-173815.

33 See for example Cara Biascucci and Robert Prentice, *Behavioral Ethics in Practice: Why We Sometimes Make the Wrong Decision* (Routledge, 2021), which highlights the interdisciplinary nature of the field.

34 Deane-Peter Baker and Nicole Townsend, 'Defiant Heroism or Wilful Disobedience? John "Jack" Hinton and the Ethics of Disobedience in War', *Defence Studies* 24, no. 3 (2024): 349–364, at: https://doi.org/10.1080/14702436.2024.2346486.

35 'MERLIN: Military Ethics Research Lab and Innovation Network', *UNSW Canberra: School of Humanities & Social Sciences* (website), at: https://www.unsw.edu.au/canberra/about-us/our-schools/hass/our-research/merlin.

36 Chunpeng Zhai, Santoso Wibowo and Lily D Li, 'The Effects of Over-Reliance on AI Dialogue Systems on Students' Cognitive Abilities: A Systematic Review', *Smart Learning Environments* 11, no. 28 (2024), at: https://doi.org/10.1186/s40561-024-00316-7.

37 Michael Mayer, 'Trusting Machine Intelligence: Artificial Intelligence and Human-Autonomy Teaming in Military Operations', *Defense & Security Analysis* 39, no. 4 (2023): 521–538, at: https://doi.org/10.1080/14751798.2023.2264070.

38 Mathias Klaus, 'Transcending Weapon Systems: The Ethical Challenges of AI in Military Decision Support Systems', *Humanitarian Law & Policy*, 24 September 2024, at: https://blogs.icrc.org/law-and-policy/2024/09/24/transcending-weapon-systems-the-ethical-challenges-of-ai-in-military-decision-support-systems.

39 Sayed Fayaz Ahmad, Heesup Han, Muhammad Mansoor Alam, Mohd. Khairul Rehmat, Muhammad Irshad, Marcelo Arraño-Muñoz and Antonio Ariza-Montes, 'Impact of Artificial Intelligence on Human Loss in Decision Making, Laziness and Safety in Education', *Humanities & Social Sciences Communications* 10 (2023), at: https://doi.org/10.1057/s41599-023-01787-8.

40 The term was introduced by psychologists Susan Fiske and Shelley Taylor in 1984 to define those whose capacity to process information is limited, so they take shortcuts whenever possible. Susan T Fiske and Shelley E Taylor, *Social Psychology: Social Cognition* (Addison-Wesley, 1984).

41 See Mary L Cummings, 'Automation and Accountability in Decision Support System Interface Design', *The Journal of Technology Studies* 32, no. 1 (2006): 23–31, at: https://scholar.lib.vt.edu/ejournals/JOTS/v32/v32n1/pdf/cummings.pdf; and Jacob K Farnsworth, Kent D Drescher, Jason A Nieuwsma and Robyn B Walser, 'The Role of Moral Emotions in Military Trauma: Implications for the Study and Treatment of Moral Injury', *Review of General Psychology* 18, no. 4 (2014): 249–262, at: https://www.researchgate.net/publication/270508829_The_Role_of_Moral_Emotions_in_Military_Trauma_Implications_for_the_Study_and_Treatment_of_Moral_Injury.

42 Shannon Vallor, 'Moral Deskilling and Upskilling in a New Machine Age: Reflections on the Ambiguous Future of Character', *Philosophy & Technology* 28 (2015): 107–124.

43 Ned Dobos, 'Two Concepts of Moral Injury: Moral Trauma and Moral Degradation', in Tom Frame (ed.), *Moral Injury: Unseen Wounds in an Age of Barbarism* (UNSW Press, 2016), pp. 126–134.

44 Molloy, *Drones in Modern Warfare*, p. 82.

45 Dominika Kunertova, 'Drones Have Boots: Learning from Russia's War in Ukraine', *Contemporary Security Policy* 44, no. 4 (2023): 576–91, at: https://www.tandfonline.com/doi/full/10.1080/13523260.2023.2262792.

46 Samuel Bendett and David Kirichenko, 'Battlefield Drones and the Accelerating Autonomous Arms Race in Ukraine', *Modern War Institute*, 10 January 2025, at: https://mwi.westpoint.edu/battlefield-drones-and-the-accelerating-autonomous-arms-race-in-ukraine.

47 For example, the AI software and systems developed by Ukraine-based companies Swarmer and ZIR System. Rebecca Pool, 'Drones with Edge AI: The Future of Warfare?', *EE Times Europe*, 5 March 2025, at: https://www.eetimes.eu/drones-with-edge-ai-the-future-of-warfare; 'Ukrainians Have Created an AI for Drones That Automatically Identifies and Strikes Targets', *The Odessa Journal*, 19 November 2024, at: https://odessa-journal.com/ukrainians-have-created-an-ai-for-drones-that-automatically-identifies-and-strikes-targets.

48 Thomas Maxwell, 'Ukraine Is Using Millions of Hours of Drone Footage to Train AI for Warfare', *Gizmodo*, 20 November 2024, at: https://gizmodo.com/ukraine-is-using-millions-of-hours-of-drone-footage-to-train-ai-for-warfare-2000541633#.

49 Nikoleta Manakitsa, George S Maraslidis, Lazaros Moysis and George F Fragulis, 'A Review of Machine Learning and Deep Learning for Object Detection, Semantic Segmentation, and Human Action Recognition in Machine and Robotic Vision', *Technologies* 12, no. 15 (2024), at: https://doi.org/10.3390/technologies12020015.

50 Matthew Krupczak, 'Manoeuvreist Doctrine in the Age of Autonomy', *Australian Army Research Centre* (website), 21 March 2024, at: https://researchcentre.army.gov.au/library/land-power-forum/manoeuvreist-doctrine-age-autonomy.

51 Rachel Huang, Jonathan Pedoeem and Cuixian Chen, 'YOLO-LITE: A Real-Time Object Detection Algorithm Optimized for Non-GPU Computers', *2018 IEEE International Conference on Big Data (Big Data)*, Seattle, 10–13 December 2018 (IEEE, 2018), pp. 2503–2510.

52 Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto and Hartwig Adam, 'MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications', *ArXiv*, abs/1704.04861 (2017).

53 Krupczak, 'Manoeuvreist Doctrine in the Age of Autonomy'.

54 Molloy, *Drones in Modern Warfare*, p. 79.

55 Tom M Mitchell, *Machine Learning* (McGraw Hill, 1997), p. 2.

56 Joy Buolamwini and Timnit Gebru, 'Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification', Proceedings of the 1st Conference on Fairness, Accountability and Transparency, *Proceedings of Machine Learning Research* 81 (2018): 77–91, at: https://proceedings.mlr.press/v81/buolamwini18a.html.

57 Patrick Grother, *Face Recognition Vendor Test (FRVT) Part 8: Summarizing Demographic Differentials*, NIST Interagency Report NIST IR 8429 ipd (U.S. Department of Commerce, July 2022), at: https://doi.org/10.6028/NIST.IR.8429.ipd.

58 Anh Nguyen, Jason Yosinski and Jeff Clune, 'Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images', *2015 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, 7–12 June 2015 (IEEE, 2015), pp. 427–36, at: https://doi.org/10.1109/CVPR.2015.7298640.

59 Olivia Farrer, 'Understanding AI Vulnerabilities', *Harvard Magazine*, 21 March 2025, at: https://www.harvardmagazine.com/2025/03/artificial-intelligence-vulnerabilities-harvard-yaron-singer; 'Groundbreaking BBC Research Shows Issues with Over Half the Answers from Artificial Intelligence (AI) Assistants', *BBC Media Centre*, 11 February 2025, at: https://www.bbc.com/mediacentre/2025/bbc-research-shows-issues-with-answers-from-artificial-intelligence-assistants.

60 Kate Devitt, Michael Gan, Jason Scholz and Robert Bolia, *A Method for Ethical AI in Defence* (Defence Science and Technology Group, February 2021), at: https://www.dst.defence.gov.au/publication/ethical-ai; see also Tara Roberson, Stephen Bornstein, Rain Liivoja, Simon Ng, Jason Scholz and Kate Devitt, 'A Method for Ethical AI in Defence: A Case Study on Developing Trustworthy Autonomous Systems', *Journal of Responsible Technology* 11 (2022): 100036, at: https://doi.org/10.1016/j.jrt.2022.100036.

61 Athena AI, at: athenadefence.ai.

62 Paul E Green and Vithala R Rao, 'Conjoint Measurement for Quantifying Judgmental Data', *Journal of Marketing Research* 8, no. 3 (1971): 355–363, at: https://doi.org/10.2307/3149575; Jordan J Louviere, 'Conjoint Analysis Modelling of Stated Preferences: A Review of Theory, Methods, Recent Developments and External Validity', *Journal of Transport Economics and Policy* 22, no. 1 (1988): 93–119, at: https://jtep.org/journal/conjoint-analysis-modelling-of-stated-preferences-a-review-of-theory-methods-recent-developments-and-external-validity.

63 'Departments of Defence and Veterans' Affairs Human Research Ethics Committee', *Australian Government: Department of Defence* (website), at: https://www.defence.gov.au/defence-activities/research-innovation/research-committees/ddva-hrec.

64 The Red Card is a pocket-sized red card listing the Australian Army's baseline conditions and steps for opening fire.

65 Danny Campbell, Marco Boeri, Edel Doherty and W George Hutchinson, 'Learning, Fatigue and Preference Formation in Discrete Choice Experiments', *Journal of Economic Behavior & Organization* 119 (2015): 345–363, at: https://doi.org/10.1016/J.JEBO.2015.08.018.

66 Jordan J Louviere, David A Hensher and Joffre D Swait, *Stated Choice Methods: Analysis and Applications* (Cambridge University Press, 2000).

67 Bhaskar Chakravorti, 'AI's Trust Problem: Twelve Persistent Risks of AI That Are Driving Scepticism', *Harvard Business Review*, 3 May 2024, at: https://hbr.org/2024/05/ais-trust-problem.

68 Tomislav Furlanis, Takayuki Kanda and Dražen Brščić, 'Robots as Moral Environments', *AI & Society* 39 (2024): 1746–1767, at: https://doi.org/10.1007/s00146-023-01656-7.

69 National Artificial Intelligence Centre, 2024, https://www.industry.gov.au/publications/ai-impact-navigator

70 Australian Army, *Robotic & Autonomous Systems Strategy v2.0* (Commonwealth of Australia, 2022), at: https://researchcentre.army.gov.au/rico/robotic-and-autonomous-systems/robotic-autonomous-systems-ras-strategy.

71 Neil C Rowe, 'The Comparative Ethics of Artificial-Intelligence Methods for Military Applications', *Frontiers in Big Data*, 12 September 2022, at: https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2022.991759/full.

72 HW Meerveld, RHA Lindelauf, EO Postma and M Postma, 'The Irresponsibility of Not Using AI in the Military', *Ethics and Information Technology* 25, no. 14 (2023), at: https://doi.org/10.1007/s10676-023-09683-0.

RESEARCHCENTRE.ARMY.GOV.AU

CG: MAY002-26