Ministry of Defence

AMBITIOUS, SAFE, RESPONSIBLE

Our approach to the delivery of AI-enabled capability in Defence

June 2022

This policy statement should be read in conjunction with the Defence AI Strategy 2022

Ambitious, Safe, Responsible 2022

v1.0

June 2022

Conditions of Release

Executive Summary

Defining Artificial Intelligence

Defence understands Artificial Intelligence (AI) as a family of general-purpose technologies, any of which may enable machines to perform tasks normally requiring human or biological intelligence, especially when the machines learn from data how to do those tasks.

The Defence AI Strategy sets out our view of the strategic opportunities and challenge presented by the emergence of AI as a transformative and disruptive new technology. **Realising the benefits of AI – and countering threats and challenges associated with the use of AI by others – is one of the most critical strategic challenges of our time.**

AI will enable our people to make powerful use of previously unimaginable quantities of data. It will improve decision-making and the delivery of operational effect. There is a strong case for the development and use of an AI system where it would be demonstrably beneficial or result in a more ethical outcome. In an era of increasing global competition and given resource limitations, it is imperative that we deliver maximum effectiveness and efficiency using AI across the spectrum of Defence activities.

We also recognise that the nature of AI gives rise to risks and concerns about possible impact on humans. These can be particularly acute in a Defence context. If we do not address them, we risk losing public consent, seeing our ability to operate undermined, and exacerbating the irresponsible behaviour of others – not to mention undermining our ability to deliver Defence capability.

We believe that a broad 'systems' perspective will ensure AI-related issues are addressed systematically and effectively. **By focusing on outcomes, delivered through clear frameworks & processes , and guided by our conviction that AI can be a powerful force for good, we will ensure that we:**

- **are ambitious in terms of the tools and operational effects we seek to deliver;**
- **enable – rather than constrain – the delivery of those tools and effects; and**
- **deliver and use AI-enabled capability in a safe and responsible manner.**

Within Defence, this will be achieved through a number of overlapping approaches. We will:

- set clear organisational intent and ambition for the adoption and exploitation of AI, backed up with defined roles and responsibilities, as set out in the Defence AI Strategy;
- continue to apply our robust **safety and regulation** regimes;
- always comply with our national and international **legal** obligations; and
- set out a clear framework and processes for ensuring **ethical adoption** of the technology.

This is a positive blueprint for effective, innovative and responsible AI adoption. Fundamentally, the best way to develop AI systems for Defence which serve our operational needs and reflect the values of those we serve is to: set ambitious direction; ensure clarity and certainty around our approaches; and provide assurance to colleagues in Defence, industry and wider society – thereby demonstrating trustworthiness.

Ambitious delivery of capability

We aspire to exploit AI comprehensively, accelerating 'best in class' AI-enabled capabilities into service in order to make all parts of Defence significantly more efficient and effective. To do this, we must ensure that we are always ambitious in the ways in which we incorporate AI into Defence capabilities where AI is the appropriate tool to adopt. We will not adopt AI for its own sake; it is not an 'end' in itself.

We intend that our approach will **enable** – rather than constrain – the adoption and exploitation of AI-enabled solutions and capabilities across Defence. We will **empower** teams developing and delivering concepts, technologies and solutions to explore ambitious ideas and use cases. We will provide them with clear frameworks to support the early identification and resolution of safety, legal and ethical risks; this will give them the confidence to explore the full potential of the technology while complying with policy and other essential requirements. We will encourage them to identify wider factors impeding their progress – such as policy or process – in the expectation that appropriate solutions will be identified and implemented rapidly.

We want to harness the creativity and innovation found across Defence and the private sector. This includes the necessary problem-solving approaches to bring those ambitious use-cases to life in an appropriate way. Since risks and challenges may exist in respect of any part of the 'system of systems' across the full lifecycle of the capability, we are clear that solutions may similarly be found across the full 'system of systems' and lifecycle.

In other words, the issue may not lie in 'what' the capability is designed to do, but 'how' it does it, and how we ensure that AI is used effectively and appropriately within it. We will ensure that suitable methods are adopted across our enterprise to 'design out' problems, and that we rigorously test AI-enabled solutions. Further information can be found in the Defence AI Strategy.

As a general enabling technology, AI has been dubbed 'the new electricity'.[1] *Clearly there are important differences. One useful point of comparison, is that initial planning to deliver a major system (ultimately to be powered by electricity) would focus on the desired outcome and level of ambition. Discussion might be mainly about vision and willingness to push boundaries, although planners would also consider possible delivery routes and likely be aware of engineering limitations.*

Initial planning would not normally hinge on discussions about whether the likely incorporation of electricity *meant that the system could be delivered safely and responsibly. These issues would be addressed systematically through design, manufacture, use in service and disposal, even if this necessitated changes in operational approaches, or additional Research & Development to provide technical solutions.*

This is similar to the way we think about AI. We start from the belief that AI is a powerful tool, and that we must be confident in our vision for AI-enabled capability. Safe and responsible outcomes are then a function of Defence-wide processes, rather than of early self-imposed limitations which would risk being arbitrary, constraining and habitually out-dated, given the speed of technological advances.

[1] Andrew Ng: *Why AI Is the New Electricity*, gsb.stanford.edu/insights/Andrew-ng-why-ai-new-electricity

Our approach and AI-enabled weapons

We will focus on outcomes, exploring ambitious options rather than filtering out ideas or concepts when they are still on the drawing board. **We do not rule out incorporating AI within weapon systems**. In practice, however, some concepts and capabilities may prove impossible to deliver in a safe and responsible manner – and we are very clear that **there must be context-appropriate human involvement in weapons which identify, select and attack targets.** This could mean some form of real-time human supervision, or control exercised through the setting of a system's operational parameters.

We believe that AI can substantially augment the performance of our people and significantly enhance our capabilities. However, given concerns about the ethics and risks of delegating certain decisions to AI, it is also important to state that **we do not believe that 'more autonomous' necessarily means 'more capable'**. We believe that Human-Machine Teaming[2] delivers the best outcomes, in terms of overall effectiveness, optimal use of resources, the practicalities of integration and the ease with which we can address issues arising; it is therefore our default approach to AI adoption.

The appropriate degree of system 'autonomy' and type of 'human control' need to be considered carefully on a case-by-case basis.

Looking at AI-enabled systems from a technological perspective: the precise degree of AI-enabled autonomy needed within any given capability or component will depend on the specific nature of the system or the concept for its use. Operational outcomes typically depend on speed and accuracy, in terms of assessment and then action. These can be enabled to a very high degree without the requirement to delegate problematic levels of discretion or judgement to the AI.

Looking at the same systems from a People perspective: we can exert satisfactory and rigorous human control over AI-enabled systems without always requiring some form of real-time human supervision. Indeed, doing so may act as an unnecessary and inappropriate constraint on operational performance. For example, to defend a maritime platform against hypersonic weapons we may need defensive systems which can detect incoming threats and open fire faster than a human could react.

Another crucial point is that all new weapons, means and methods of warfare are subject to a rigorous review process for compliance with International Humanitarian Law and other applicable international law. Determinations as to the necessary scope and application of context-appropriate human involvement will be done similarly systematically. We also adjust our operating procedures to ensure that we stay within the boundaries of the law that applies at the time.

This approach is reflected in our policy with regards to international debates on 'Lethal Autonomous Weapon Systems, set out in more detail at **Annex C**.

[2] Joint Concept Note 1/18 – Human Machine Teaming

Key challenges to Defence AI Adoption

The use of AI in a Defence context raises a number of interlinked issues and challenges, which development teams and users will need to take into consideration. These include:

- **Algorithmic Bias**: the risk that biased datasets used to train AI systems could result in discriminatory outcomes and disproportionate harms for certain groups of users;

- **Responsibility & Accountability**: the need to ensure that delegation of tasks or decisions to AI systems does not lead to a 'responsibility gap' between systems that take decisions or make recommendations, and the human commanders responsible for them;

- **Unpredictability**: the risk that some AI systems may behave unpredictably, particularly in new or complex environments, or as they learn and adapt over time;

- **Unintended Consequences and Incentives**: the potential for AI-enabled systems to have unintended side-effects on human behaviour, through enabling certain incentives, or influencing other systems beyond their intended effect;

- **People Implications**: the need to think differently about what is expected of people, and the impact of AI on people, as AI-enabled systems create opportunities to automate 'dull, dirty or dangerous' tasks; and

- **Human Control**: when using AI-enabled systems for Defence purposes, the need to understand the appropriate form of human involvement required for any given application or context.

These issues may be encountered individually or in combination. Some create safety issues, requiring us to ensure that our AI does not cause inadvertent harm or danger as a part of its use. Some give rise to ethical issues, requiring us to ensure that our use of the technology aligns with our values, and those of the society we represent. Some are important elements in ensuring that the particular capability complies with our domestic and international legal obligations (e.g. relating to data protection, privacy or International Humanitarian Law).

In handling these issues properly, we must maintain the trust and goodwill of our key stakeholders - including our service personnel - allies, and partners in the private sector. Without this, we risk slowing innovation, losing important opportunities to collaborate – and reduced public consent for the use of these technologies.

By adopting a safe and responsible approach to AI, we also have an opportunity to set a positive example to others, encouraging the safe and responsible use of AI globally.

To achieve these outcomes, we are establishing a clear framework which will **provide support and clarity to the teams within Defence and beyond** who are developing and operating our AI-enabled systems.

Using AI Safely

A number of the key challenges associated with adopting AI for Defence pose particular issues for safety.

The **unpredictability of some AI systems**, particularly when applied to new and challenging environments, increase the risks that unforeseen issues may arise with their use. The relative difficulties with **interpreting how some forms of AI systems learn and make decisions** present new challenges for the testing, evaluation and certification of such systems. In addition, **the high potential impact of AI-enabled** systems for Defence raises the stakes for potential side effects or unintended consequences, particularly when they could cause harms for those interacting with them.

Broadly, **this is not a new challenge for the Department**. Defence is bound by UK law and has a robust regime for compliance. Defence activities also include those that are inherently dangerous and require additional risk management beyond that of our statutory obligations.

Where Defence has certain derogations, exemptions or disapplication's from UK legislation and regulations, it is the Department's policy and practice to maintain arrangements that produce outcomes that are, so far as practicable, at least as good as those required by UK legislation. This is reflected in **Health, Safety and Environmental Protection (HS&EP) in Defence – Policy statement by the Secretary of State for Defence**.

A strict compliance with safety rules is therefore essential to Defence's use of any new technology.

The Defence Safety Authority (DSA) contributes to Defence capability, reputation and effectiveness through the setting, and enforcement of Defence Regulations for Health, Safety and Environmental Protection and supports the Ministry of Defence by providing independent, evidence-based assurance.

The DSA conducts horizon scanning activity with regards to the development in AI capability. The DSA will continue to set Defence Regulation and conduct enforcement activity across in-service capability and will examine how AI capability can be assured in future.

Within the DSA, the Defence Accident Investigation Branch (DAIB) provides Defence with an accident and incident investigation capability conducting impartial and expert no-blame safety investigations across all domains, with a focus on the identification and understanding of all accident factors. This may include accidents related to AI capability.

Using AI Legally

Defence's activities are governed by a range of legislative provisions which ensure our work is undertaken in accordance with the law. This legislation protects fundamental freedoms and human rights, while giving the MOD the powers it needs to keep citizens safe and secure in the modern world

Defence always seeks to abide by its legal obligations across the full range of activities from employment law, to privacy and procurement, and the law of armed conflict, also known as International Humanitarian Law (IHL). It has robust practices and processes in place to ensure its activities and its people abide by the law. These practices and processes are being – and will continue to be – applied to AI-enabled capabilities.

Deployment of AI-enabled capabilities in armed conflict needs to comply fully with IHL, satisfying the four core principles of distinction, necessity, humanity and proportionality. **We are very clear that use of any system or weapon which does not satisfy these fundamental principles would constitute a breach of international law.**[3]

Article 36 legal reviews ensure that commanders, service personnel, politicians, the UK public and our allies can be assured that UK weapons are lawful. Additional Protocol 1 of the Geneva Convention, requires 'in the study, development, acquisition or adoption of a new weapon, means or method of warfare . . . to determine whether its employment would, in some or all circumstances, be prohibited by [Additional Protocol I] or by any other [applicable] rule of international law'.

As with any new and emerging technology, the Ministry of Defence is therefore conscious of the need to be aware of any legal issues that may arise with the use of AI in Defence, whether as a means or method of warfare or in a 'back office' system, and has robust review processes in place.

Our development and use of AI technologies will always be in accordance with the body of applicable UK and international law.

[3] For an explanation of how these legal and our ethical principles apply to discussions around Lethal Autonomous Weapon Systems see Annex C.

Using AI Ethically

The MOD is dedicated to the protection of UK people, territories, values and interests at home and overseas. We must be ethical – and be seen to ethical – in our AI development and use to protect UK values and retain the trust and support of our citizens, key stakeholders, allies and partners. This includes recognising the potential for cases where AI could be an important tool helping us to promote ethical outcomes and where it might be unethical not to use AI.

We have therefore developed **ethical principles for AI in Defence (Annex A)**, designed to lead our overall approach to AI-enabled technologies and systems across the full range of possible use cases, from back office to decision support and battlespace capabilities.[4] They will apply to our development and use of AI across their lifecycle, and also to *AI-enabled systems*: platforms, processes or systems of which AI forms some part. These principles will;

- steer our approach to the key challenges of AI, in particular providing direction around responsible development, training and use (see implementation section);
- form the core of the UK's approach to creating agreed norms for AI in Defence internationally, working with partners and allies to shape the global development of AI in the direction of freedom, openness and democracy; and
- provide characteristics for AI systems which teams across Defence will be expected to follow in the development of new AI-enabled systems.

Setting out the principles is a key step towards ensuring a responsible and safe approach to the technology. Effective implementation and evolution of the principles will be required to ensure that our development of AI matches our requirement for safe and responsible innovation.

Partnerships and Consultation

To develop these Principles, the MOD worked in partnership with the Centre for Data Ethics and Innovation (CDEI). The first in the world of its kind, the CDEI leads the UK Government's work to enable trustworthy innovation using data and AI. Guided by an advisory body of internationally-recognised experts, the Centre works with partners across the public sector, industry and academia, in the UK and internationally, to identify and tackle barriers to responsible innovation. The MOD Principles are the result of over 18 months of consultation with over 100 expert stakeholders from around the world.[5]

As part of the consultation process, the MOD has convened an **AI Ethics Advisory Panel**, a group of experts in computer science, AI ethics and military ethics. This Panel ensured MOD benefitted from specialist insight and challenge, and played a crucial role supporting the development of the Principles. Further details on its role and current membership are available at **Annex B**.

Going forward into implementation of these approaches, our commitment to transparency and consultation with industry and allies will continue. We will continue to work with the CDEI as this area evolves to ensure that our ongoing approach matches responsible best practice. We will also continue to engage proactively with leading experts from academia and industry, including the Ethics Advisory Panel.

[4] For an explanation of how these ethical principles and also the legal principles of International Humanitarian Law apply to discussions around Lethal Autonomous Weapon Systems, see Annex C.

[5] Our joint methodology comprised desk and interview-based research exploring key issues, including the unique ethical challenges posed by AI in Defence. We deployed a range of methods to test and iterate the principles, including one-to-one and group interviews, roundtables and workshops with experts from across Defence, industry, academia, civil society, law, government and frontline military personnel. We also tested the principles against a range of hypothetical Defence AI use cases, and against a typical Defence AI system lifecycle. Further details on the process will be outlined in a joint MoD-CDEI report which we intend to publish in 2022.

Governance

This policy statement should be read in conjunction with the Defence AI Strategy 2021, which provides more detail across the breadth of AI issues and activities relevant to Defence.

In line with the governance model set out in the Strategy:

- the **2nd Permanent Secretary (2nd PUS) and the Vice Chief of the Defence Staff** will **oversee and drive AI-related activity** across Defence;

- **2nd PUS has specific responsibility for AI policy** across Defence, including oversight of our ethical framework and responsibility for taking forward measures within the delegated model to ensure effective implementation;

- the **Permanent Secretary (PUS) remains responsible for health, safety and environmental protection,** supported by the Chief Operating Officer (COO) and the Director Health Safety & Environmental Protection;

- 2nd PUS (and other senior officials as required) will be supported (in terms of policy development) by the **Defence AI & Autonomy Unit** (DAU), part of Defence Science & Technology (DST), and advised by the **AI Ethics Advisory Panel**. Scientific and technical advice will be provided by the MOD Chief Scientific Adviser (supported by DST and the Defence Science & Technology Laboratory (Dstl)) and the Chief Information Officer (supported by Defence Digital). Independent technical advice and review will be provided by the Defence Science Expert Committee (DSEC);

- **Top Level Budget (TLB) duty holders and trading fund agency chief executives are senior duty holders for safety** and are responsible for designating the duty holders in their organisation who manage activities which could be a risk to life. Each **TLB** organisation will have an **accountable officer responsible for AI Ethics** implementation.

Implementation – building justified trust

Having set out our overall approach to adopting AI ambitiously, safely and responsibly in this document, effective implementation will be critical. We must maximise the benefit we extract from AI while also demonstrating trustworthiness, both in terms of the breadth of our portfolio of AI-enabled tools and capabilities and the specifics of individual use cases. These two goals are linked. **Ensuring our use of AI is safe, reliable and responsible doesn't impede innovation; it's key to collaboration and ensuring systems deliver the outcomes we need.**

Our approach will therefore be:

- **outward-facing:** we will be transparent, engaging consistently with and welcoming challenge from industry, academia, civil society, international partners and allies;

- **applied across the entire AI system lifecycle:** Defence teams will be able to articulate how a safe and responsible approach is being followed across an AI system's full lifecycle, covering all Defence Lines of Development;

- **context-specific:** the vast range of potential use cases for AI across Defence means that application of policy approaches cannot be uniform or technology-specific, but must take into consideration the particular requirements of each project.

Annex A: Ethical Principles for AI in Defence

Preamble: Our intent for the ethical use of AI in Defence

The MOD is committed to developing and deploying AI-enabled systems responsibly, in ways that build trust and consensus, setting international standards for the ethical use of AI for Defence. The MOD will develop and deploy AI-enabled systems for purposes that are demonstrably beneficial: driving operational improvements, supporting the Defence Purpose, and upholding human rights and democratic values.

The MOD's existing obligations under UK law and international law, including as applicable international humanitarian law (IHL) and international human rights law, act as a foundation for Defence's development, deployment and operation of AI-enabled systems. These ethical principles do not affect or supersede existing legal obligations. Instead, they set out an ethical framework which will guide Defence's approach to adopting AI, in line with rigorous existing codes of conduct and regulations.

These principles are applicable across the full spectrum of use cases for AI in Defence, from battlespace to back office, and across the entire lifecycle of these systems.

First principle: Human-Centricity

The impact of AI-enabled systems on humans must be assessed and considered, for a full range of effects both positive and negative across the entire system lifecycle.

Whether they are MOD personnel, civilians, or targets of military action, humans interacting with or affected by AI-enabled systems for Defence must be treated with respect. This means assessing and carefully considering the effects on humans of AI-enabled systems, taking full account of human diversity, and ensuring those effects are as positive as possible. These effects should prioritise human life and wellbeing, as well as wider concerns for human kind such as environmental impacts, while taking account of military necessity. This applies across all uses of AI-enabled systems, from the back office to the battlefield.

The choice to develop and deploy AI systems is an ethical one, which must be taken with human implications in mind. It may be unethical to use certain systems where negative human impacts outweigh the benefits. Conversely, there may be a strong ethical case for the development and use of an AI system where it would be demonstrably beneficial or result in a more ethical outcome.

Second principle: Responsibility

Human responsibility for AI-enabled systems must be clearly established, ensuring accountability for their outcomes, with clearly defined means by which human control is exercised throughout their lifecycles.

The increased speed, complexity and automation of AI-enabled systems may complicate our understanding of pre-existing concepts of human control, responsibility and accountability. This may occur through the sorting and filtering of information presented to decision-makers, the automation of previously human-led processes, or processes by which AI-enabled systems learn and evolve after their initial deployment. Nevertheless, as unique moral agents, humans must always be responsible for the ethical use of AI in Defence.

Human responsibility for the use of AI-enabled systems in Defence must be underpinned by a clear and consistent articulation of the means by which **human control** is exercised, and the nature and limitations of that control. While the level of human control will vary according to the context and capabilities of each AI-enabled system, the ability to exercise human judgement over their outcomes is essential.

Irrespective of the use case, **Responsibility** for each element of an AI-enabled system, and an articulation of risk ownership, must be clearly defined from development, through deployment – including redeployment in new contexts – to decommissioning. This includes cases where systems are complex amalgamations of AI and non-AI components, from multiple different suppliers. In this way, certain *aspects* of responsibility may reach beyond the team deploying a particular system, to other functions within the MOD, or beyond, to the third parties which build or integrate AI-enabled systems for Defence.

Collectively, these articulations of human control, responsibility and risk ownership must enable clear **accountability** for the outcomes of any AI-enabled system in Defence. There must be no deployment or use without clear lines of responsibility and accountability, which should not be accepted by the designated duty holder unless they are satisfied that they can exercise control commensurate with the various risks.

Third principle: Understanding

AI-enabled systems, and their outputs, must be appropriately understood by relevant individuals, with mechanisms to enable this understanding made an explicit part of system design.

Effective and ethical decision-making in Defence, from the frontline of combat to back-office operations, is always underpinned by appropriate understanding of context by those making decisions. Defence personnel must have an appropriate, context-specific understanding of the AI-enabled systems they operate and work alongside.

This level of understanding will naturally differ depending on the knowledge required to act ethically in a given role and with a given system. It may include an understanding of the general characteristics, benefits and limitations of AI systems. It may require knowledge of a system's purposes and correct environment for use, including scenarios where a system should not be deployed or used. It may also demand an understanding of system performance and potential fail states. Our people must be suitably trained and competent to operate or understand these tools.

To enable this understanding, we must be able to verify that our AI-enabled systems work as intended. While the 'black box' nature of some machine learning systems means that they are difficult to fully explain, we must be able to audit either the systems or their outputs to a level that satisfies those who are duly and formally responsible and accountable. Mechanisms to interpret and understand our systems must be a crucial and explicit part of system design across the entire lifecycle.

This requirement for context-specific understanding based on technically understandable systems must also reach beyond the MOD, to commercial suppliers, allied forces and civilians. Whilst absolute transparency as to the workings of each AI-enabled system is neither desirable nor practicable, public consent and collaboration depend on context-specific shared understanding. What our systems do, how we intend to use them, and our processes for ensuring beneficial outcomes result from their use should be as transparent as possible, within the necessary constraints of the national security context.

Fourth principle: Bias and Harm Mitigation

Those responsible for AI-enabled systems must proactively mitigate the risk of unexpected or unintended biases or harms resulting from these systems, whether through their original rollout, or as they learn, change or are redeployed.

AI-enabled systems offer significant benefits for Defence. However, the use of AI-enabled systems may also cause harms (beyond those already accepted under existing ethical and legal frameworks) to those using them or affected by their deployment. These may range from harms caused by a lack of suitable privacy for personal data, to unintended military harms due to system unpredictability. Such harms may change over time as systems learn and evolve, or as they are deployed beyond their original setting. Of particular concern is the risk of discriminatory outcomes resulting from algorithmic bias or skewed data sets. Defence must ensure that its AI-enabled systems do not result in unfair bias or discrimination, in line with the MOD's ongoing strategies for diversity and inclusion.

A principle of bias and harm mitigation requires the assessment and, wherever possible, the mitigation of these biases or harms. This includes addressing bias in algorithmic decision-making, carefully curating and managing datasets, setting safeguards and performance thresholds throughout the system lifecycle, managing environmental effects, and applying strict development criteria for new systems, or existing systems being applied to a new context.

Fifth principle: Reliability

AI-enabled systems must be demonstrably reliable, robust and secure.

The MOD's AI-enabled systems must be suitably reliable; they must fulfil their intended design and deployment criteria and perform as expected, within acceptable performance parameters. Those parameters must be regularly reviewed and tested for reliability to be assured on an ongoing basis, particularly as AI-enabled systems learn and evolve over time, or are deployed in new contexts.

Given Defence's unique operational context and the challenges of the information environment, this principle also requires AI-enabled systems to be secure, and a robust approach to cybersecurity, data protection and privacy.

MOD personnel working with or alongside AI-enabled systems can build trust in those systems by ensuring that they have a suitable level of understanding of the performance and parameters of those systems, as articulated in the principle of understanding.

Annex B: The Ministry of Defence AI Ethics Advisory Panel

The MOD AI Ethics Advisory panel is an informal advisory board to the 2nd Permanent Secretary for Defence in his role as senior responsible owner for AI Ethics in the department.

The panel's purpose is to convene a combination of expert voices from Defence, academia, industry and civil society to advise the 2nd Permanent Secretary on the development of policy relating to safe and responsible development and use of AI.

The panel is advisory only, and has no formal decision-making powers, but will be responsible for scrutinising the MOD's ongoing approach to responsible and ethical AI. Panellists were appointed by the MOD on the basis of their expertise across the subjects of AI development, AI ethics, military ethics and international law.

As of the date of publication, the current panel has met three times, and has served a key role in providing scrutiny and advice on the crafting of the ethical principles and potential methods of implementation. The panel has not been involved in the creation of policy related to Lethal Autonomous Weapons Systems, nor the department's policy on AI safety.

The current panel membership is as follows:

Laurence Lee, 2nd Permanent Secretary for Defence (*Chair*)

Professor Dapo Akande, Director of the Oxford Institute for Ethics, Law and Armed Conflict

Professor Nick Colosimo, Global Engineering Fellow & Technologist, BAE systems and Visiting Professor, Cranfield University (Centre for Autonomous & Cyber-Physical Systems).

Dr Merel Ekelhof, Foreign Exchange Officer at the US DoD Joint AI Center and former Lead Researcher on AI and Autonomy at UNIDIR, attending the panel in her personal capacity.

Tabitha Goldstaub, Founder of CognitionX and chair of the AI Council

Dr Darrell Jaya-Ratnam, Managing Director, DIEM Analytics

Professor Peter Lee, Professor of Applied Ethics, University of Portsmouth

Professor Dame Angela McLean, Chief Scientific Advisor at the Ministry of Defence

Richard Moyes, Managing Director and co-founder, Article 36

Professor Gopal Ramchurn, Director, UKRI Trustworthy Autonomous Systems Hub and the University of Southampton

Polly Scully, Director for Strategy at the Ministry of Defence

Professor Mariarosaria Taddeo, Associate Professor and Senior Research Fellow, Oxford Internet Institute, University of Oxford; Dstl Ethics Fellow, Alan Turing Institute.

Lt Gen Roly Walker, Deputy Chief of the Defence Staff for Military Strategy and Operations

Professor David Whetham, Professor of Ethics and the Military Profession, Kings College London

Dominic Wilson, Director General for Security Policy, Ministry of Defence

ANNEX C: Lethal Autonomous Weapon Systems (LAWS)

One of the most significant concerns around the use of AI in Defence is around introducing elements of autonomy to the use of weapons systems. This subject has already been the source of significant international debate, particularly under the UN's Group of Government Experts on LAWS. The following section sets out the UK's position on this potential use case for AI-enabled systems.

AI can enable systems – including weapons – to exhibit some measure of autonomy: deciding and acting to accomplish desired goals, within defined parameters, based on acquired knowledge and an evolving situational awareness. This potentially could lead to weapons that identify, select and attack targets without context-appropriate human involvement. That is not acceptable – the United Kingdom does not possess fully autonomous weapon systems and has no intention of developing them.

We strongly believe that AI within weapon systems can and must be used lawfully and ethically. Sharing the concerns of Governments and AI experts around the world, we therefore oppose the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifecycle. The use of such weapons could not satisfy fundamental principles of International Humanitarian Law, nor our own values and standards as expressed in our AI Ethical Principles. Human responsibility and accountability cannot be removed – irrespective of the level of AI or autonomy in a system. The UK will always clearly establish authorities, thus human responsibility, and accountabilities whenever UK forces deploy weapon systems which incorporate AI.

We will continue to work closely with international allies and partners to address the opportunities and risks around autonomy in weapons systems. Global governance for such systems is a difficult task. It will be challenging to reach international agreement on definitions for full or partial autonomy on a technical or systems level. It is also important to ensure any approach allows for rapid technological advancement, and doesn't become redundant or isn't able to be circumvented as technology develops. Such international processes must be inclusive, and involve all key actors in this space if they are to be effective.

We believe the best approach is to focus on building norms of use and positive obligations to demonstrate how degrees of autonomy in weapons systems can be used in accordance with international humanitarian law – with suitable levels of human control, accountability and responsibility. Setting out those characteristics that would make it inherently impossible for a system to comply with international humanitarian law is key to this, and we will continue to engage actively in the international arena to reach consensus on them. The UN Group of Government Experts on LAWS under the Convention for Certain Conventional Weapons will continue to be our primary avenue for such discussions. Our own approach, driven by the AI Ethical principles, is to build understanding, best practice and codes of conduct through which we can achieve ethical outcomes in our use of AI.