

Issue

Brief

ISSUE NO. 570
AUGUST 2022



Managing Expectations: Explainable A.I. and its Military Implications

Shimona Mohan

Abstract

The rapid uptake of artificial intelligence (AI) in the military in the past couple decades has been accompanied by a slow but gradual build-up in attempts to understand how these AI systems work to achieve better results in military operations. The idea behind what is called 'eXplainable AI' (XAI), and the technologies driving it, are a manifestation of this trend. The question, however, is if XAI in its current form is the solution that it is expected to be. Modelled as a scoping exercise, this brief seeks to cover each of these aspects and explore the wider implications of the use of XAI in the military, including its development, deployment, and governance.

The potential for military use has often been the driving force of technological innovation around the world. In recent years, there has been a notable increase in the development and deployment of highly advanced disruptive technologies for defence purposes, and artificial intelligence (AI) has become the poster child for this trend. Only a few years ago, the current gamut of applications of AI in military operations would have been dismissed as fodder for fiction. Today, with advances in emerging technologies in the area of lethal autonomous weapons systems (LAWS) and the continuous integration of AI and machine learning (ML) into the back-end of existing military computing systems, military applications of AI systems around the world are only set to increase in number and intensity.¹ This surge is accompanied by new ideas of ensuring that the deployed military AI systems are more compatible with human use and have smaller margins of error. One such idea is the development of what is called eXplainable AI (XAI), i.e. AI and ML systems that make it possible for human users to understand, appropriately trust, and effectively manage AI.²

This brief explains why such systems are a necessity in the military, what XAI is and how it functions, examples of where and how it has been applied so far, and evaluates its use and regulation. The brief uses both primary and secondary research sources, including interviews with expert stakeholders from different geographies and disciplines, either currently or formerly from government, defence services, civil society, and academia. It aims to analyse the current status of XAI in the military, and pave the way for more targeted research.

eXplaining the Need for XAI

The dual-use characteristic of AI ensures that any upgrades in its development and deployment for civilian purposes can also be applied to their military counterparts, and vice versa. For instance, facial recognition softwares that unlock mobile phones and help automatically tag friends in pictures posted on social media have been employed by the Israeli army to find and track Palestinian military objectives.³ The same softwares are also being used by the Ukrainian defence ministry to identify potentially undercover or deceased Russian soldiers.⁴ Elsewhere, AI-based computer vision programmes for self-driving cars like Tesla have been used by Azerbaijan to navigate autonomous unmanned aerial vehicles (UAVs) in the Nagorno-Karabakh conflict.⁵ In the United States, algorithms like those that customise what-to-watch lists for individual users on streaming platforms are projected to become part of the cognitive equipment of the armed forces to advise soldiers in communications-denied or resources-constrained environments.⁶

Indeed, from countries with well-established military AI ecosystems such as France,⁷ to ones that are emerging in this domain like India,⁸ countries across the globe are investing millions of dollars in AI. These states recognise AI as a force multiplier in military operations⁹ that grants its users an upper-hand against rivals by contextually processing large amounts of data; identifying trends, patterns, people and objects of interest; piloting systems and processes in both critical and non-critical military functions; and predicting and recommending courses of action that may help in aiding, or in some cases even replacing, human decision-making in high-stakes, time-sensitive situations.

AI also has the potential for defensive military use, and can be trained to recognise, flag, and neutralise malicious software.¹⁰ Operationally, AI can enable military teams to maintain or expand warfighting capacity without the need to increase or upskill personnel strength, both of which require time and additional recurring costs that militaries may not be able or willing to spare.¹¹

eXplaining the Need for XAI

Despite its purported advantages, however, military uses of AI are not without significant challenges. AI systems can be purposefully programmed to cause death or destruction, either by the users themselves or through an attack on the system by an adversary. Unintended harm can also result from inevitable margins of error which can exist or occur even after rigorous testing and proofing of the AI system according to applicable guidelines. Indeed, even ‘regular’ operations of deployed AI systems are mired with faults that are only discoverable at the output stage.¹² At that point, the result of such oversights could already be irreversible and may cause irreparable damage to military operations in terms of compromising personnel, equipment, and/or information.

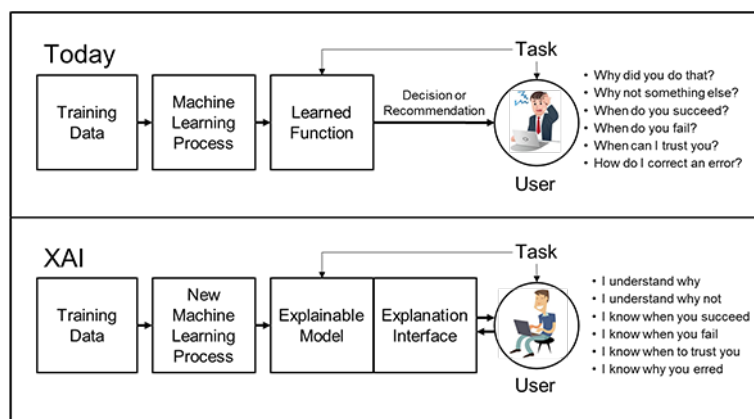
A primary cause for such faults is flawed training datasets and commands, which can result in misrepresentation of critical information as well as unintended biases. Another, and perhaps far more challenging, reason is issues with algorithms within the system which are undetectable and inexplicable to the user. As a result, AI has been known to produce outputs based on spurious correlations and information processing that does not follow the expected rules, similar to what is referred to in psychology as the ‘Clever Hans effect’.^{a,13}

A study tracking publicly available instances of bias in 133 AI systems across industries from 1988 to 2021 discovered that 44.2 percent (59 systems) demonstrated a gender bias, and 25.7 percent (34 systems) exhibited both gender and racial biases, due to the quality of the datasets used to train these systems as well as their algorithms.¹⁴ Transplanted into a conflict environment, deploying such AI systems could mean, for instance, that a woman of a race against which this programme is biased, who thus belongs to the intersecting demographic of race and gender that is doubly plagued with biases, could be misidentified by the computer vision or facial recognition software of an autonomous weapon system (AWS) as a non-human object. Therefore, even if the AWS was designed to not engage with humans, the system could perceive it to be in accordance with its programmed rules to engage with her and possibly cause her injury or death since it does not consider her as human. Such inaccuracies result from what is referred to as the ‘black box’ of AI, i.e. the algorithmic processing within AI and ML systems that is not explainable to or understandable by human operators.¹⁵

a Consider an ML-based system that classified images of horses as such not based on common features of the animal, but due to similar watermarks and copyright tags on the images of horses that the system was fed. See: Ian Sample, “Computer Says No: Why Making AIs Fair, Accountable and Transparent is Crucial,” *The Guardian*, November 5, 2017, <https://www.theguardian.com/science/2017/nov/05/computer-says-no-why-making-ais-fair-accountable-and-transparent-is-crucial>

In a bid to counter the black box problem and the inadvertent harm it could cause if applied in military operations, the Defense Advanced Research Projects Agency (DARPA) under the US Department of Defense (DoD) began a multi-project programme in 2016 to develop ‘white box’ or ‘glass box’ XAI to ensure the explainability and understandability of AI systems and bring in effective human-machine interactions (see Fig. 1).¹⁶ While the term ‘XAI’ is relatively new, the idea is not: the problem of ‘explainability’ has existed since the mid-1970s when technical researchers studied explanations for expert systems.¹⁷ Following DARPA’s announcement, interest around XAI grew and a niche research ecosystem began to emerge to understand how dual-use AI applications could be made more explainable and understandable.

Fig. 1:



Source: DARPA¹⁸

While there was no consolidated definition or characterisation given of what XAI is—computing models, best practices of AI applications, or a mix of both—researchers from both technical and non-technical areas dived into developing or interpreting XAI in their own way. Software professionals attempted to design AI- and ML-based systems to test the explainability of the system to the human end-user, and observed that most methods to attain XAI tended to fall into one of two broad categories: ante-hoc and post-hoc.¹⁹

Ante-hoc methods involve using basic models that can execute low-level AI functions. They ensure that AI systems are intrinsically explainable throughout the processing of information because of the simplicity of their processing and decision-making. Meanwhile, post-hoc methods are those that assign transparency to the output received from another underlying black box model in their own separate model of processing. These methods can be model agnostic, i.e. applied to any type of model that is treated as a black box; or model specific, i.e. limited to operating on certain types of models that correspond with a certain model logic.

Although both methods ensure a threshold of explainability, ante-hoc methods were found to be highly limited in their application potential and scalability in real-world instances. It was observed that higher-performing AI systems remain less explainable and understandable due to the more complex processing that generates their outputs, and the more explainable systems are, by default, low-performing due to their simplistic parameters and information processing.

Yet, it is incorrect to assume that the higher-performing AI systems would result in more accurate or unbiased decisions than the low-performing ones. The measure of the performance level of the AI system is directly proportional to how hard it would be to detect its flaws and understand the reasons behind why it processed information in a certain way. This performance-explainability trade-off has remained a cardinal problem of XAI, and the development of newer models and methods continue to face the challenge of unravelling it.

eXploring the Applications of XAI

The incumbent continuous development of XAI is envisioned to lead the advent of ‘third wave’ AI systems,²⁰ where AI-enabled machines understand the context and environment in which they operate, and build their own underlying explanatory models to characterise real-world phenomena. While research around XAI first emerged due to military interests of the US and the country has remained the forerunner in this domain, civilian uses of XAI have also emerged in other parts of the world in the past five years. These have given global legitimacy to the idea of explainable and understandable AI systems being an advantageous innovation on various fronts.

By 2019, Big Tech companies like Google,²¹ Microsoft,²² and IBM²³ had introduced their respective XAI toolkits. Medical service providers such as Fujitsu Laboratories and Hokkaido University in Japan announced the development of a new technology based on XAI that offers customised suggestions to patients about their health by identifying interlinkages from their past medical data.²⁴ In financial services, a project on developing a standard for XAI led by China’s first digital bank, WeBank, was approved by the Institute of Electrical and Electronics Engineers (IEEE) in 2020, and marked the world’s first industry standard that specifically solves the black box problem in AI applications in banking.²⁵

Such applications of XAI prove the ubiquitous nature of AI systems and their related functionalities, and highlight how technology introduced primarily for military purposes can have cross-cutting civilian utility. Moreover, the comparatively low-stakes functions of civilian applications of XAI can render its adoption into civilian architectures faster than military ones, ultimately providing an inadvertent testing ground for some of its use-case characteristics before potential military deployment.²⁶

Given today’s defence landscape, a number of countries have initiated projects to incorporate AI into their defence architecture. For example, the US plans to spend almost a billion dollars to develop hundreds of new military AI projects in 2022,²⁷ China is aiming to actively ‘intelligentise’ its warfare tactics using AI,²⁸ Russia is progressively developing 150 AI-enabled military systems,²⁹ and Israel is working to systematically incorporate AI across all systems in its military.³⁰ The Indian government stated this year as well that it plans to

eXploring the Applications of XAI

develop 25 defence-specific AI products by 2024.³¹ While the development and deployment of XAI may seem to be a recent issue requiring a far more in-depth analysis of the intersection between defence technology and AI ethics, some countries have already publicly announced their intention to do so.

In 2020, the Swedish Defence Research Agency initiated research on XAI use in the military with the following multi-pronged goal: a) supporting military end-users to create mental models of how AI systems function; b) allowing specialists to gain insight and extract knowledge from the hidden tactical and strategic behaviour of AI systems; c) helping developers to identify flaws or bugs and address them prior to a system's deployment; and d) more effectively obey the rules of war.³²

In the same year, the Indian Navy aboard the INS Valsura began to explore the possibility of developing and applying XAI in maritime operations through a call for papers. They reasoned that there is often a gap in trust when using AI technologies due to the black box, and if decision-makers do not trust the decisions recommended by AI in defence services, the top brass would be reluctant to adopt the technology.³³

The US DoD has also forayed further into XAI this year and solicited applications to develop XAI models for command-and-control decision aids to use in Multi-Domain Operation (MDO) wargaming.³⁴ Most recently, in June 2022, the Ministry of Defence (MoD) of the United Kingdom (UK) released a policy paper³⁵ in conjunction with their *Defence AI Strategy 2022*, highlighting principles for the ethical use of AI in defence. One of the principles, titled 'understanding', underlines how the black box problem makes defence-related AI systems difficult to explain, but mechanisms to interpret and understand the systems must be a crucial and explicit part of system design across their entire lifecycle. This paper considers previous work by the UK Information Commissioner's Office (ICO) and the Alan Turing Institute in collaboration with Project ExplAIIn—the trio had earlier published a document on XAI in 2020 to provide practical advice on explaining decisions made by AI systems, in a manner that meets legal requirements as well as technical and governance best practice.³⁶

As is clear, many countries have started engaging more concretely with XAI for defence purposes, but there are as yet no examples of monitoring and analysis of any progress indicators. The efficacy of their commitment to and operationalisation of XAI in the military can only be ascertained over a longer term of a few years.

X AI is clearly a subject of great interest within the AI paradigm, but is the fascination around it justified?

While XAI undoubtedly has its advantages, its development and deployment also warrants legitimate concerns. Foremost amongst these relates to its performance-explainability trade-off: if explainability, which is the cardinal property of XAI and the foundation of other approaches to ethical AI, is inverse to performance of AI/ML systems, will politicians and legislators see merit in continuing research and investments into XAI? Renewed R&D into XAI may eventually lead to a breakthrough wherein effective XAI models can be integrated into high-performing AI systems, but there is no indication as to a timeline for such a development.

Even if such a breakthrough occurs, would it be enough? Some experts believe there may always be explainability lacunae in XAI systems, since every explanation provided by the systems overshadows another potential explanation.^{b,37} Thus, the aspiration should not be a blanket accomplishment of high-performing AI systems with effective XAI models, but a more simplistic, albeit also harder to achieve, XAI model that provides contextual clarity about the information processing of the AI system—who would need to know what, and how should the XAI model present that to them. A suggested method to ensure this is explainability by design (EbD): designing a framework of required explanations, generation of explanations along this framework by the XAI model, and a machine learning-supported evaluation of the output by the end user to ensure better future compliance with the framework.³⁸

Apart from attempting to fix the performance-explainability trade-off in XAI models, it is essential to consider the human halves of the operation since they often become the weakest links in tech-enabled command chains. Monitoring and evaluation of the original DARPA projects show that user cognitive load to interpret explanations provided by the XAI models can hinder user performance.³⁹ This means that even if XAI models explain their processes to the human operator, the latter would need to be well-versed with dynamic and

b Thinking back to the Clever Hans reference, why is a picture of a horse classified as a horse? Because of the tail? Or hooves? Or background of the picture? Any number or combination of explanations could eclipse others, and current XAI systems cannot discern which one holds value to the user.

often non-uniform models. Understandability of the AI functioning through appropriate training has to be ensured for the end user of the XAI model, otherwise, the explainability of the system would serve no purpose. Since even humans cannot justify all their actions, specific principles need to be in place to regulate the user uptake of the explanations provided by XAI.⁴⁰


Principles should also be developed through research for XAI systems, ultimately leading to the development of standards and licenses for AI use in their separate domains.⁴¹ For instance, the European Union (EU) approved a six-year-long project in 2019 on the prospect and details of the civilian application of XAI, focusing on how to design transparency in ML models, produce controlled black-box explanations, and formulate ethical and legal standards for AI.⁴² There is also an appetite from within civil society to develop responsible AI solutions that are explainable, although the groundwork for this is being laid out solely for civilian applications.

In terms of application-agnostic frameworks, researchers from Google, Microsoft and IBM created two Responsible AI Licenses (RAIL) in 2019,⁴³ which constitute the world's first independent, accredited certification programme for responsible AI. The licenses test AI/ML systems on the five categories of explainability, fairness, accountability, robustness, and data quality to ascertain their status as responsible AI applications. In a similar move, the Responsible AI Institute released a beta version of their RAII Certification in 2022, which is developed under the World Economic Forum's (WEF) Global AI Action Alliance (GAIA).⁴⁴ The RAII Certification is finalised through a robust process of risk-benefit analyses, calibrating assessment frameworks, validating findings with experts, and testing and evaluation of results by an independent council.⁴⁵

Proponents of military XAI should push to replicate similar standards and licenses for integration in the military AI architecture to ensure that problems of explainability, reliability, and bias within their AI systems are institutionally resolved, regardless of whether or not there is political will to do so in individual cases. For example, several countries are unwilling to be obligated to conduct standardised legal reviews of LAWS, preferring a more voluntary and/or internal review that relies on their own parameters.⁴⁶ While military XAI standards or certifications would be a significant first step and a contemporaneous solution, it would not be enough in the long term. Fragmented development of XAI and related responsible AI frameworks by singular stakeholders like private companies and military contractors could encompass the potential for vested interests and problems with subsequent standardisation. There needs to be independent oversight and an effective governance mechanism for XAI in the military that does not treat it as simply another tick-box for military AI development.⁴⁷

The development and use of XAI, especially in military applications, is still nascent and holds neither clarity on how it is being developed across the board nor a standardised approach towards regulating what already has been or what is in the pipeline to be developed and deployed. There are two key takeaways in this analysis of XAI in the military.

It is now becoming increasingly common to see technology outpacing the law, and more so in areas such as military AI which are pioneers of motivated technological innovation. While XAI would generally be seen as a positive value-laden technology, at the end of the day, it is just another tool in use by the military, and has its inherent strengths and weaknesses like any other.⁴⁸ As such, it needs to evolve within a governance framework that guides its effective and ethical use, and prevents it from being vulnerable to potential misuse, inefficacy or redundancy.

Research on XAI still seems to be fragmented, sectoral, and mostly clustered in certain geopolitical spaces, which creates knowledge silos and limits the dissemination of new ideas. Moreover, there is a lot of idealism around XAI in the military specifically,⁴⁹ and therefore critical analyses and balanced research are required to provide a holistic perspective of its prospects. Future engagements with this issue domain should seek to bridge these gaps and generate more interdisciplinary and inter-sector analyses, while also attempting to gather and include viewpoints from various contexts. 

Shimona Mohan is an emerging scholar with research interests at the nexus of technology and security policy.

- 1 Forrest E. Morgan et al, *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*, RAND Corporation, 2020, https://www.rand.org/pubs/research_reports/RR3139-1.html#:~:text=A%20steady%20increase%20in%20the,mitigate%20the%20most%2Dextreme%20risks
- 2 Jessica Newman, “Explainability Won’t Save AI,” *The Brookings Institution*, May 19, 2021, <https://www.brookings.edu/techstream/explainability-wont-save-ai/>
- 3 Emma Roth, “The Israeli Army is Using Facial Recognition to Track Palestinians, Former Soldiers Reveal,” *The Verge*, November 8, 2021, <https://www.theverge.com/2021/11/8/22769933/israeli-army-facial-recognition-palestinians-track>
- 4 Paresh Dave and Jeffrey Dastin, “Ukraine Has Started Using Clearview AI’s Facial Recognition During War,” *Reuters*, March 15, 2022, <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearview-ais-facial-recognition-during-war-2022-03-13/>
- 5 Robyn Dixon, “Azerbaijan’s Drones Owned the Battlefield in Nagorno-Karabakh — and Showed Future of Warfare,” *The Washington Post*, November 11, 2020, https://www.washingtonpost.com/world/europe/nagorno-karabakh-drones-azerbaijan-aremenia/2020/11/11/441bcbd2-193d-11eb-8bda-814ca56e138b_story.html
- 6 Brandon Knapp, “The Army Wants to Give Soldiers a Netflix-Like Recommendation on the Battlefield,” *C4ISRNet*, May 3, 2018, <https://www.c4isrnet.com/it-networks/2018/05/02/the-army-wants-to-give-soldiers-a-netflix-like-recommendation-on-the-battlefield/>
- 7 Julian Turner, *Intelligent Design: Inside France’s €1.5Bn AI Strategy - Global Defence Technology | Yearbook 2018*, Global Defense Technology – NRI Digital, 2018, https://defence.nridigital.com/global_defence_technology_yearbook_2018/intelligent_design_inside_frances_15bn_ai_strategy
- 8 Joe Saballa, “India ‘Increasingly Focusing’ on AI for Military Applications”. *The Defense Post*, February 14, 2022, <https://www.thedefensepost.com/2022/02/14/india-ai-military/>
- 9 Greg Hadley, “Former Google CEO: AI Will Be ‘Force Multiplier Like You’ve Never Seen Before’,” *Air Force Magazine*, March 4, 2022, <https://www.airforcemag.com/former-google-ceo-ai-force-multiplier-pentagon/#:~:text=Schmidt’s%20intense%20enthusiasm%20for%20artificial,of%20that%20an%20AI%20assistant.%E2%80%9D>
- 10 Forrest E. Morgan et al, *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*
- 11 Forrest E. Morgan et al, *Military Applications of Artificial Intelligence: Ethical Concerns in an Uncertain World*

- 12 Kelsey Atherton, “Understanding the Errors Introduced by Military AI Applications,” *The Brookings Institution*, May 6, 2022, <https://www.brookings.edu/techstream/understanding-the-errors-introduced-by-military-ai-applications/>
- 13 Eugen Lindwurm, “Deep Learning, Meet Clever Hans,” *Medium*, August 15, 2020, <https://towardsdatascience.com/deep-learning-meet-clever-hans-3576144dc5a9>
- 14 Genevieve Smith and Ishita Rustagi, “When Good Algorithms Go Sexist: Why and How to Advance AI Gender Equity,” *Stanford Social Innovation Review*, March 31, 2021, https://ssir.org/articles/entry/when_good_algorithms_go_sexist_why_and_how_to_advance_ai_gender_equity
- 15 Arthur Holland, *The Black Box, Unlocked: Predictability and Understandability in Military AI*, United Nations Institute for Disarmament Research (UNIDIR), 2020, <https://unidir.org/sites/default/files/2020-09/BlackBoxUnlocked.pdf>
- 16 Matt Turek, “Explainable Artificial Intelligence (XAI),” *Defense Advanced Research Projects Agency (DARPA)*, <https://www.darpa.mil/program/explainable-artificial-intelligence>
- 17 Amina Adadi and Mohammed Berrada, “Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI),” *Institute of Electrical and Electronics Engineers (IEEE)*, 2018, <https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=8466590>
- 18 Turek, “Explainable Artificial Intelligence (XAI)”
- 19 Alejandro Barredo Arrieta *et al.*, “Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI,” *Information Fusion (Volume 58) – ScienceDirect*, (2020), <https://www.sciencedirect.com/science/article/abs/pii/S1566253519308103>
- 20 Rajesh Uppal, “DARPA (AIE, XAI and AI Next) Developing ‘Third Wave’ AI Based Adaptive Military Systems That Are Trustworthy, Learn Continuously, and Explain Their Rationale,” *International Defense, Security & Technology Inc.*, May 5, 2021, <https://idstch.com/technology/ict/darpa-aie-xai-and-ai-next-developing-third-wave-ai-based-adaptive-military-systems-that-are-trustworthy-learn-continuously-and-explain-their-rationale/>
- 21 Tracey Frey, “Google Cloud AI Explanations to Increase Fairness, Responsibility, and Trust,” *Google Cloud Blog*, November 21, 2019, <https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-ai-explanations-to-increase-fairness-responsibility-and-trust>
- 22 “Fairlearn,” Microsoft Fairlearn, 2019, <https://fairlearn.org/>
- 23 “Explainable AI,” IBM, 2019, <https://www.ibm.com/watson/explainable-ai>
- 24 Mark Bowen, “Fujitsu and Hokkaido University Develop ‘Explainable AI’ Technology,” *Intelligent CIO APAC*, February 4, 2021, <https://www.intelligentcio.com/apac/2021/02/04/fujitsu-and-hokkaido-university-develop-explainable-ai-technology/>

- 25 Cheng Yu, "Project to Set Standard on Explainable AI Approved," *China Daily*, July 29, 2020, <https://global.chinadaily.com.cn/a/202007/29/WS5f213362a31083481725cec1.html>
- 26 Interviewed technology and security expert, Center for a New American Security (USA), June 2022
- 27 John Keller, "Pentagon to Spend \$874 Million on Artificial Intelligence (AI) and Machine Learning Technologies Next Year," *Military Aerospace Electronics*, June 4, 2021, <https://www.militaryaerospace.com/computers/article/14204595/artificial-intelligence-ai-dod-budget-machine-learning>
- 28 Ryan Fedasiuk, "We Spent a Year Investigating What the Chinese Army is Buying. Here's What We Learned," *Politico*, November 10, 2021, <https://www.politico.com/news/magazine/2021/11/10/chinese-army-ai-defense-contracts-520445>
- 29 "Artificial Intelligence and Autonomy in Russia, Issue 41" *Centre for Naval Analyses (CNA)*, June 27, 2022, <https://www.cna.org/centers/cna/sppp/rsp/russia-ai>
- 30 Seth Frantzman, "Israel Unveils Artificial Intelligence Strategy For Armed Forces," *Defense News*, February 11, 2022, <https://www.defensenews.com/artificial-intelligence/2022/02/11/israel-unveils-artificial-intelligence-strategy-for-armed-forces/>
- 31 Frantzman , "Israel Unveils Artificial Intelligence Strategy For Armed Forces"
- 32 Swedish Defence Research Agency (FOI), *Explaining Artificial Intelligence with XAI*, 2020, <https://www.foi.se/en/foi/news-and-pressroom/news/2020-10-05-explaining-artificial-intelligence-with-xai.html>
- 33 The Indian Navy, *Webinar on AI For Data Driven Navy*, 2020, <https://www.indiannavy.nic.in/insvalsura/sites/default/files/AI%20Webinar.pdf>
- 34 Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) Programs, *Explainable AI for Complex Decision Making for Command and Control in Multi-Domain Operations (MDO)*, 2022, <https://www.sbir.gov/node/2214713>
- 35 Ministry of Defence, Government of the United Kingdom, *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence*, 2022, <https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence#annex-a-ethical-principles-for-ai-in-defence>
- 36 Arnav Joshi, "Moving Forward on Explainable AI - New Guidance from the UK ICO and Turing Institute," *Clifford Chance*, June 11, 2020, <https://www.cliffordchance.com/briefings/2020/06/moving-forward-on-explainable-ai---new-guidance-from-the-uk-ico-.html>

- 37 Interviewed responsible AI expert, Responsible AI Institute (USA), June 2022
- 38 Trung Dong Huynh *et al*, “Explainability-By-Design: A Methodology to Support Explanations in Decision-Making Systems,” *ArXiv*, 2022, <https://arxiv.org/ftp/arxiv/papers/2206/2206.06251.pdf>
- 39 David Gunning, “DARPA’s Explainable AI (XAI) Program: A Retrospective,” *Applied AI Letters – Wiley Online Library* (2021), <https://onlinelibrary.wiley.com/doi/full/10.1002/ail2.61>
- 40 Interviewed AI policy expert, AI Asia Pacific Institute (Singapore/UAE), June 2022
- 41 Interviewed AI policy expert, AI Asia Pacific Institute
- 42 Community Research and Development Information Service (CORDIS), *European Commission, Science and Technology for the Explanation of AI Decision Making*, 2022, <https://cordis.europa.eu/project/id/834756>
- 43 “Responsible AI Licenses V0.1,” Responsible AI Licenses (RAIL), 2019, <https://www.licenses.ai/ai-licenses/>
- 44 “RAI | RAII Certification,” Responsible AI Institute (RAII), 2022, <https://www.responsible.ai/certification>
- 45 Interviewed responsible AI expert, Responsible AI Institute
- 46 Vincent Boulanin, “Implementing Article 36 Reviews in Light of Increasing Autonomy of Weapons,” *Stockholm International Peace Research Institute (SIPRI)*, 2015, <https://www.sipri.org/sites/default/files/files/insight/SIPRIInsight1501.pdf>
- 47 Interviewed AI ethics expert, University of Pretoria (South Africa), June 2022
- 48 Nick Starck, David Bierbrauer and Paul Maxwell, “Artificial Intelligence, Real Risks: Understanding – and Mitigating – Vulnerabilities in the Use of AI,” *Modern War Institute*, January 18, 2022, <https://mwi.usma.edu/artificial-intelligence-real-risks-understanding-and-mitigating-vulnerabilities-in-the-military-use-of-ai/>
- 49 Interviewed technology and security expert, Center for a New American Security



Ideas . Forums . Leadership . Impact

20, Rouse Avenue Institutional Area,
New Delhi - 110 002, INDIA
Ph. : +91-11-35332000. Fax : +91-11-35332005
E-mail: contactus@orfonline.org
Website: www.orfonline.org