On Explainable AI Solutions for Targeting in Cyber Military Operations

Clara Maathuis
Open University of the Netherlands, Heerlen, The Netherlands
clara.maathuis@ou.nl

Abstract: Nowadays, it is hard to recall a domain, system, or problem that does not use, embed, or could be tackled through AI. From early stages of its development, its techniques and technologies were successfully implemented by military forces for different purposes in distinct military operations. Since cyberspace represents the last officially recognized operational battlefield, it also offers a direct virtual setting for implementing AI solutions for military operations conducted inside or through it. However, planning and conducting AI-based cyber military operations are actions still in the beginning of development. Thus, both practitioner and academic dedication is required since the impact of their use could have significant consequences which requires that the output of such intelligent solutions is explainable to the engineers developing them and also to their users e.g., military decision makers. Hence, this article starts by discussing the meaning of explainable AI in the context of targeting in military cyber operations, continues by analyzing the challenges of embedding AI solutions (e.g., intelligent cyber weapons) in different targeting phases, and is structuring them in corresponding taxonomies packaged in a design framework. It does that by crossing the targeting process focusing on target development, capability analysis, and target engagement. Moreover, this research argues that especially in such operations carried out in silence and at incredible speed, it is of major importance that the military forces involved are aware of the following. First, the decisions taken by the intelligent systems embedded. Second, are not only aware, but also able to interpret the results obtained from the AI solutions in a proper, effective, and efficient way. From there, this research draws possible technological and human-oriented methods that facilitate the successful implementation of XAI solutions for targeting in military cyber operations.

Keywords: cyber operations, cyber weapons, military operations, targeting, artificial intelligence, explainable AI

1. Introduction

> *"From this place, and from this day forth, begins a new era in the history of the world, and you can all say that you were present at its birth."* (Goethe)

Crossing centuries through the classical OODA (Observe, Orient, Decide, Act) loop, a long series of wars were conducted since the 17th century, spanning the local and world wars, and going into todays and future's wars. Accordingly, the Observe topic moved from telescope (wars in the 17th century) to radio and radar (WWII) and goes to network (future wars). The Orient topic moved from weeks (wars in the 17th century) to hours (WWII) and is going to be continuous (future wars). The Decide topic moved from months (wars in the 17th century) to days (WWII) and will be immediate (future wars). The Act topic transformed to according to season (wars in the 17th century), to weeks (WWII), and goes to minutes (future wars) (Lehto, 2016). These developments were possible due to significant technological advancements that the scientific community and industry professionals proposed and continue to do. Among these developments AI finds its place. No matter if used for planning optimization, target identification, or effects assessment, AI showed impressive results in different prediction, simulation, or exploration problems (Samek & Müller, 2019) tackled in cyberspace. Here, the operations are conducted using cyber weapons/capabilities to achieve military objectives inside and/or outside it (Maathuis, Pieters & Van den Berg, 2018a) having (un)foreseen (in)direct effects on targets and collateral entities (Maathuis, Pieters & Van den Berg., 2018b). Compared to other military operations, cyber military operations take place in silence and at high speed, allow early intelligence gathering and preparations, and imply diverse options for building cyber weapons corresponding to targets' vulnerabilities, nature, aim, and context. In these moments, the decisions made and implications considered, have to be clear and understandable to the stakeholders involved.

Centrally located in the ongoing developments and successes of AI is machine learning, and in particular deep learning. While the improvements, accessibility, and use of these paradigms represent a mindset change and a turnover on data focused applications, also imply understanding and dealing with the decisions made and the results obtained by models: a complex and difficult task. However, recently, a starting point in this direction is the DARPA program (DARPA, 2016) aimed at creating human understandable AI models through a set of design options considering the explainability-performance trade-off. This research line captured the interest of academic researchers and practitioners from different fields (Adadi & Berrada, 2018), but a theory for XAI and a

universally agreed definition is lacking (Samek & Müller, 2019). Moreover, especially in this domain, for targeting is important that decision makers are aware which and how data is used since data could be scarce in this battlefield or could come from several battlefields, understand which decisions the model takes inside and outside cyberspace, and are conscious of the results proposed in order to integrate and deploy it properly with minimal or without risks in the operational field since such risks and their corresponding effects could be experienced both digitally and physically and at large scale due to cyberspace's particularities like interconnectivity and dynamic nature. Given the aspects abovementioned, understanding and applying XAI in cyber military operations is a complex task and has to be tackled considering multiple angles. On this matter, to the best of our knowledge XAI was not defined and tackled in the military cyber context, thus we aim to address this having the following objectives:

- To propose a definition and common understanding for XAI in military cyber operations.
- To address challenges of XAI in military cyber operations through techno-military and socio-ethical lenses focusing on the first phases of the targeting process.
- To raise the level of awareness and responsibility of decision makers on the design and implementation of XAI models for military decision support and broader in the military cyber domain.
- To contribute to the design of strategies, standards, and methods for XAI in the military cyber domain. Further, to stress the need for education in XAI for both current and future decision makers.

To achieve these objectives, multidisciplinary research is conducted based on extensive literature review and analysis. Furthermore, the contributions of this research are twofold. First, to merge and highlight various aspects that should be considered when grasping, developing, and evaluating XAI with a focus on three phases of targeting in military cyber operations. And second, to serve as a design framework with concrete recommendations for developments for decision makers.

The remainder of this article is structured as follows. Section 2 discusses related research concerning theoretical and practical aspects of XAI. Section 3 tackles the targeting process while reflecting on the phases that this research focuses on. Section 4 addresses the necessity of proposing XAI models, reflects on different types of participating stakeholders, analyses characteristics of good explanations, and advances a definition for XAI in this domain. Section 5 considers different types of explanations, explanation methods, and evaluation criteria and mechanisms for explanation methods. Section 6 tackles challenges for XAI for targeting in military cyber operations. Section 7 reflects on the findings of this research and discusses future ideas.

2. Related Research

The general interest in developing AI techniques significantly grew in the last decade, and this is of particular interest in respect to proposing both theoretical and practical approaches for XAI in different domains.

Concerning the meaning and development of XAI, Arrieta et al. (2020) summarize previous efforts and consider that XAI is "widely acknowledged as a crucial feature for the practical deployment of AI models" and position the audience of such models being a key element when understanding it. Accordingly, one of the reflection points is the audience and stakeholders involved in XAI. Preece et al. (2018) attribute the fact that there is no consensus for explainability and interpretability to the fact that different stakeholder communities have to deal with them, and further depict where this perspective overlap and where not. Moreover, Samek, Wiegand & Muller (2017) identify as reasons for needing XAI the following: system verification, learning from the system, and compliance to legislation. Gerlings, Shollo & Constantiou (2020) use socio-technical lenses for analyzing technical and governance aspects for XAI implementation e.g., compliance with regulation and GDPR plus bias and misinterpretation minimization. Hereof, Mohseni, Zarei & Ragan (2021) analyze different explanations and what should be explained by models further advancing a series of methods e.g., explanation satisfaction measured using criteria like user satisfaction and explanation usefulness. Thusly, Vilone & Longo (2020) provide an extensive review on explanation types, explanation methods, evaluation strategies, and future ideas for research like human-in-the-loop approach and interactive interfacing. As initiator of XAI, DARPA (2016) built a successful program finalized in 2018 with different teams working on technical aspects in the design, implementation, and evaluation of XAI models, and a team working on finding, defining, and applying psychology theories of explanation. Furthermore, the vision of the European Commission is discussed by Hamon, Junklewitz & Sanchez (2020) regarding transparency of AI models: i) documenting the AI processing chain using the technical principles of the model plus the data representations used for its design, ii) reliability of AI models that relates to their capacity of avoiding failure or malfunction due to edge cases or malicious intentions; and iii) data

protection in models for preserving security and managing risks through technical and organizational controls. Additionally, humans have the right to obtain an explanation as stressed in Recital 71 of the GDPR (Hamon, Junklewitz & Sanchez, 2020). Moreover, the U.S. DoD adopted a series of ethical principles for AI applied in (non)combat functions for upholding legal, and ethical, and policy: responsible, equitable, traceable, reliable, and governable (U.S. DoD Board, 2019).

As applications, Streich et al. (2020) analyze XAI's potential for tackling technological issues regarding producing sustainable agriculture systems corresponding to the UN sustainable development goals, and Shukla, Fan & Jennions (2020) generate guidelines for building XAI for Integrated Vehicle Health Management clearly understood by human experts using information for health assessment of the subsystems and their aircraft effects. In the medical domain, Holzinger et al. (2017) tackle the possibility of developing methods that reenact the machine decision-making process taken by XAI models in different processes. In the industry sectors, Guo (2020) argues that explainability enables trust i.e., a critical quality for 6G technology since is managing a wide range of mission critical services like autonomous driving, and Lai et al. (2020) implement a XAI model with the underlying physics of high-energy particle collisions using information encoded in the energy-momentum four-vectors of the final state particles using GANs.

In the military domain, Bistron & Piotrowski (2021) discuss applications embedding AI techniques and their impact on security sensing and show that military applications are among the ones responsible for AI development. Furthermore, Preece et al. (2019) map ISR requirements in multi-domain operations with the need of building XAI models to produce robust human-machine decision making with examples from urban terrain analysis and enhanced asset interoperability, and Hepenstal & McNeish (2020) argue that when designing military and security XAI solutions, careful consideration on the context and nature of the problem being modelled and the humans involved should be taken. In particular, Maathuis, Pieters & Van den Berg (2020) propose a multi-layered fuzzy XAI model for effects assessment and decision support for targeting in military cyber operations, and Keneni, et al. (2019) present a XAI model that depicts UAVs decisions' logic in a predefined mission.

These resources capture important aspects that should be considered when defining and proposing AI solutions in the military cyber domain but are not tailored to it and this is necessary for building trustable, explainable, and accountable military intelligent systems designed, implemented, and used in the execution of cyber military operations whose effects may by their nature cross geographical or digital boundaries. It is then the aim of this article to tackle this knowledge gap and propose a definition and model for XAI in the military cyber domain, reflect on its corresponding challenges, and further discuss possible solutions.

3. Military Targeting

To fight adversaries and achieve goals, cyber military operations are conducted by influencing their target(s) (embedding ICT elements or direct cyber targets) in several ways e.g., disrupt communication processes or alter the behavior of an audience. In this process, particularities and challenges like the dual-use nature, connectivity, dynamism, uncertainty, and attribution issue existing in cyberspace should be considered when building AI-based solutions and represent a direct reason for developing XAI-based solutions.

At the core of conducting cyber military operations is the military targeting process and corresponding Rules of Engagement (RoE) that establish the circumstances and limitations for engagement in each operation. The targeting process is defined by (NATO, 2016; U.S. Army, 2018) as selecting and prioritizing targets and matching appropriate response to them while considering operational requirements and capabilities. This process contains six phases as depicted in Figure 1 and below summarized (NATO, 2016; U.S. Army, 2018; Boothby & Schmitt, 2012):

- Phase I (Commander's intent, objectives, and guidance): define clear objectives and under which circumstances and actions these objectives should be achieved, political and strategic guidance is provided, further operational tasks are created, and targets are nominated.
- Phase II (Target development): eligible targets are analysed, vetted, validated, and prioritized resulting in a prioritised target list. Herein are also included collateral damage estimation and intelligence gain/loss assessment which implies that commanders might not engage a target in order to negatively influence the process but would benefit from the results during target selection.

- Phase III (Capability analysis): the targets are analysed and matched with capabilities to produce the desired effects while minimizing collateral damage i.e., proportionality assessment and further CoA (Course of Action) development.
- Phase IV (Commander's decision, force planning, and assignment): based on the results obtained, further assignments to specific forces are done for planning and execution considering any relevant constraints and restraints.
- Phase V (Mission planning and force execution): the mission is further planned and executed at tactical level while a final target PID (Positive Identification) is conducted with other information checks and collateral damage avoidance or minimization.
- Phase VI (Assessment): the effects produced are evaluated next to the achievement of the objectives defined, further contributing to broader assessments, input for other operations, and lessons learned.

Figure 1: Military targeting process

To go through these phases, vast amounts of procedures, methods, and models are applied merging diverse sources and types of data i.e., system, process, and human. No matter if data are further processed and used to build simulation, prediction, or combined intelligent models, models' decisions, output, and possible impact should be properly understood and justified by military decision makers. It is in this context that XAI is further defined and analysed considering its challenges and possible solutions.

4. Defining Explainable AI

Although it sounds as one of the major topics in the field of AI in the last decade, XAI finds its roots five decades before. Since then, the importance of explainablity and interpretability of intelligent systems it actually increased (Hansen & Rieger, 2019) and returned as a focus point due to the increased developments in machine learning, in deep learning which although proved its performance in a diverse plethora of tasks, is still intrinsically unable to explain its decisions in a human understandable way. In other words, as Preece, et al. (2018) stress: "it is not a new problem, nor was it ever considered a solved problem". But why XAI in military cyber operations? More concretely, what are the reasons and objectives of XAI? The following reasons are considered:

First, military cyber operations that use intelligent systems could have (in)direct, (un)intended, and (un)expected effects not only on their targets, but also on collateral assets. Especially in regards with the unintended effects, properly *understanding* where a target is localized physically and digitally, and which of its vulnerabilities is more prone to exploit in an early operational state, what is the intelligence gain/loss, and how is selected and prepared for engagement (Phase II-III) could facilitate the successful deployment of an intelligent cyber weapon (XAI based) in Phase V.

Second, coupled with understanding why and how an AI model is implemented and used, is important to *comprehend the decisions* made by a model and further *communicate* them to the systems, processes, and people involved. In particular, if a proper match to a cyber weapon is done in Phase III, its choices should be accordingly argued and communicated for execution plans and engagement from Phases IV-V.

Third, since targeting is an ongoing process carried out by multiple teams, is important that the functional mechanisms and decisions made by the AI models used (e.g., for target localization, cyber weapon selection and, effects estimation) are *justifiable* and transparent to facilitate and strengthening *control*, *trust, and accountability* (Adadi & Berrada, 2018; Burkart & Huber, 2021) along the entities involved. Moreover, this allows direct correspondence and mapping to legal checks, strategic goals, and socio-ethical values.

And the list of reasons could continue depending on the aim, activities, plus the granularity considered and appraise the uncertainty, dynamism, and strong interconnectivity aspects that characterize cyberspace. However, as these reasons reflect, XAI is first about understanding while maintaining a high level of performance (Gunning & Aha, 2019), secondly about sustaining proper communication in a human friendly way, and thirdly about justification, trust, and accountability in the systems used. As Russell & Norvig (2021) consider: "explanations are not decisions, they are stories about decisions". Thusly, we take the stance of (Molnar, 2020) that humans "desire to find meaning in the world", and after assessing that the word 'explainable' is not defined in Oxford (Oxford Dictionary, 2021) and Cambridge (Cambridge Dictionary, 2021) dictionaries, we see that the closest term is 'explanation' meaning a reason, a fact, or an excuse for something. Besides, for defining XAI in military cyber operations, we align with the definition perspectives of (DARPA, 2016; Arrieta et al., 2020; Doran et al., 2017) and consider XAI as:

XAI in the cyber military domain = a sub-field of AI that deals with the design, development, and use of methods, techniques, and technologies that provide reasons and facts for the functional mechanism, decisions, and results made by AI systems embedded in different cyber military systems and processes.

This definition could be reduced to a series of (cyber) agents that interact with each other and make use of a series of tools to produce actions. In Figure 2 we illustrate this definition while explaining its components:

Agents: entities participating in the design, development, and use of XAI in cyber military operations. Seeing their context, nature, role, and involvement in this process, they can be classified as (DARPA, 2016; Meske et al., 2021; Arrieta et al., 2020; Preece et al., 2018; Hepenstal & McNeish, 2020; Hamon et al., 2020):

- *Stakeholders*: entities involved either i) in the process of design, standardization, and certification of the models i.e., military-legal and AI regulators which should also integrate ethical aspects, or ii) in the process of theorizing, designing, developing, debugging, validating and upgrading the models i.e. AI and military cyber engineers, or iii) in the process of design, use, and assuring compliance of AI systems in and outside the organization i.e. AI mangers and cyber military decision makers.
- *Audience*: entities directly involved in the processes done by stakeholders i.e., are participatory audience or end-users i.e., they represent the other audience.

Tools: methods, techniques, and technologies used by agents to design, develop, and use XAI in cyber military operations (DARPA, 2016):

- *XAI model* is the developed AI model by stakeholders and (perhaps) participatory audience.
- *XAI interface* is the interface that generates explanations i.e., actions to the audience. Accordingly, it answers to questions like why? how? and when? e.g. Why this or why not something else? How is it done or how to tackle or correct errors? When is it successful, when not, or when will fail?

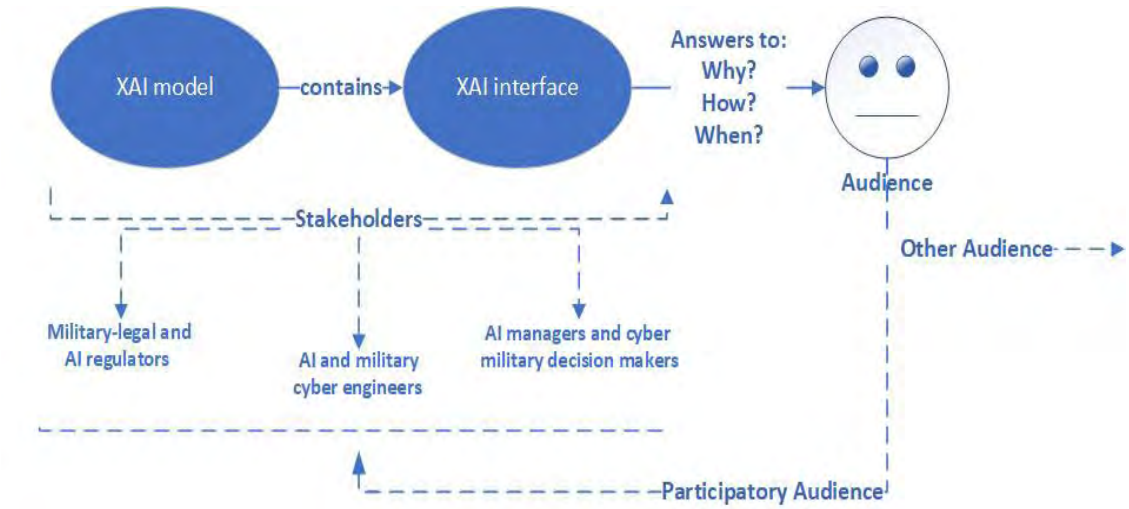

Figure 2: XAI in cyber military targeting, based on (DARPA, 2016; Meske et al., 2021)

To assess proportionality for in-house implemented military cyber operation, all stakeholders cooperate for building a XAI cyber weapon for assessing the expected collateral damage and military advantage in a human understandable way through a direct or developed XAI interface (depending on the technique) and provide a decision, meaning is this engagement proportional or disproportional, understandable, and transparent to the audience i.e., military decision makers. As seen above, a plausible question would then be: Why is proportional/disproportional? In case that the decision to engage a target is negative, then the intelligence gain/loss assessment benefits from clear explanations regarding target's location, vulnerabilities, and connections i.e., information beneficial during the ongoing or future operation(s). Then, a relevant question would be: What information could be further used?

5. Explainable AI Methods and their Evaluation

When analysing explanations regarding the role of the involved stakeholders and audience, one could consider the intention of the explanation method that refers to which question is answered, and the intention of the one(s) that use the explanation meaning how and for what should be this explanation used (Samek & Müller, 2019). Furthermore, Islam, Eberle & Ghafoor (2020) consider that explanations should be expressive, translucent, portable, accurate, have fidelity, consistent, stable, comprehensible, and contain a degree of importance. Hence, the following criteria are considered to classify the existing methods for XAI (Vilone & Longo, 2020; Samek & Müller, 2019; Arrieta et al., 2020; Kolbasin, 2018):

- *XAI model* is the developed AI model by stakeholders and (perhaps) participatory audience.
- *Problem type:* reflects the problem studied e.g., classification or regression.
- *Scope:* the objective of an explanation: *i)* local which implies that each inference of a model is explained e.g., a specific rule for proportionality assessment, while *ii)* global implies that the complete inferential process of a model is transparent and understandable as a whole e.g., general mechanism for proportionality assessment.
- *Stage:* the moment when explanations are generated: i) *ante-hoc* methods imply explainability from the beginning and during the training phase using concrete examples e.g. vulnerabilities that a target has or concrete rules that determines if engaging a target is disproportional or not, while ii) *post-hoc* methods imply mimicking model's behaviour using an external explainer during the testing phase e.g. one that would explain target's engagement with a cyber weapon or proportionality assessment rationality. These methods could be either a) *model agnostic* applied to any type of model, the model is treated as a black box, and the explanations are generated without the inspection of internal parameters, and b) *model specific* limited to specific types of models embedding specific model logic.
- *Input data:* depending on aspects like availability, volume, and data type (e.g., numerical, categorical), the explanations are considered.
- *Output format:* depending on the format of desired results (numerical, rules, textual, visual, or mixed), the explanations are considered.
- *Moment of providing explanations:* i) b*efore building the model* using methods like exploratory data analysis using different visualization methods, understanding distributions, or features analysis e.g. multiple entry points could be analysed for a series of targets to find out their centre of gravity and build

a corresponding cyber weapon in Phases III-IV, ii) d*uring building the model* directly for models like decision trees or linear ones e.g. the assessment of proportionality could be explicitly provided with its corresponding rules and results in Phases IV-V, and iii) a*fter building the model* that explain the model, its outcome, and internal processes for models based on neural networks and SVM e.g. in case of network analysis for possible target inspection done in Phase II.

However, one needs to make sure that the actions of AI models do not only meet their objectives, but actually human objectives (Russell, 2019) to consider them beneficial and trustable. These facts should be verified in both cases where AI models have a supportive or a decisive role (Samek & Müller, 2019). Accordingly, the quality of explanations depends on how the stakeholder/audience perceives them depending on their background, goals, expectations, context etc. (Atakishiyev et al., 2020). Hence, based on (Swartout & Moore, 1993; Walsh et al., 2021), the following criteria are considered to evaluate explanations of AI models:

- *Measure of suitability, sufficiency, and fidelity:* the explanations provided should be suitable, sufficient, and representative for system's actions.
- *Measure of effectiveness:* the explanations provided should be clear, useful, and understandable for the stakeholders and audience of the system.
- *Measure of performance:* the explanations provided should improve the current perceptions, believes, and perhaps competences that stakeholders and audience have on the system while making sure that this would not overhead or slow down the system.

6. Challenges of Explainable AI

As any kind of disruptive technology, XAI, presents a series of challenges in the military cyber domain. For structuring and characterizing them, socio-technical lenses are considered to capture both technical and socio-ethical elements. Thereupon, the following challenges are considered:

- Insufficient (training) data (Svenmarck et al., 2018): insufficient or not representative (training) data has a direct impact on the results provided by a model which could have significant negative consequences e.g., wrong target considered for engagement.
- Data ownership, protection, sharing, availability, and quality (Maxwell, 2020; Stahl, 2021): not only that data should be in the hand of its owners which have the duty to protect or share when and to who is necessary but should be available in different stages and maintain its quality during the processing and modelling phases.
- Loss of debuggability and transparency in development and testing (Google, 2020).
- Performance, maintenance, and robustness costs (Core et al., 2006; Gunning & Aha, 2019; Walsh et al., 2021).
- Lack of control (Google, 2020) for both stakeholders and audience.
- AI security attacks (Svenmarck et al., 2018; Morgan et al., 2020): since AI models are software-based systems, they are vulnerable to well determined adversaries that could conduct e.g., poisoning, adversarial, or differential attacks. For instance, by altering training data used for target engagement, another object could be engaged and/or the produced collateral damage could be higher than expected in Phases III-V.
- Real-live settings mirroring environment (Stahl, 2021): used for evaluating and simulating the model in realistic conditions that capture actual aspects using different simulation settings like test beds and digital twin solutions.
- Alignment with strategic and operational needs (Walsh et al., 2021): since conducting cyber military operations is a set of processes carried out by members of teams with different expertise, an alignment between their objectives and approach could be sometimes challenging.
- Alignment with commercial developments and standards (Hoadley & Lucas, 2018; Walsh et al., 2021) e.g., industry related cyber systems and solutions.
- Integration with existing systems and emerging capabilities (Walsh et al., 2021) from other military domains.
- Need for developing corresponding international political, legal, and military initiatives, standards, strategies, and methods, and further compliance with them (Deeks, 2019; Hoadley & Lucas, 2018; Samek & Müller, 2019; Morgan et al., 2020).
- Invest in education, building human capacity, and R&D (Alonso, 2020; Morgan et al., 2020).
- Fostering and building a digital ecosystem for AI (Morgan et al., 2020).

- Creation of International competition, countries asymmetries, and power imbalance (Hoadley & Lucas, 2018).
- Negative impact on human values, health, environment etc.
- Uncertainty due to the dynamism of cyberspace and the unpredictability character of conflict (Maathuis, Pieters & Van den Berg, 2016).

7. Conclusion

The mantra of AI continues to change from the more developing intelligent systems the better (Russell, 2019) to the more useful, explainable, and responsible developing intelligent systems the better. This is a long path for the scientific and industry communities from all societal domains. Of special interest, considering the potential of AI to increase the likelihood of war, to escalate ongoing conflicts, and to proliferate to malicious actors (Morgan et al., 2020) plus the new operational cyber tools that have the potential of digitally engaging their adversaries in silence and at incredible speed, it is the duty of decision makers involved in developing and/or conducting such operations, to make sure that AI is properly understood, rightly used, and that the positive aspects are promoted while taking into account, mitigating, or avoiding the negative ones (Russell, S., & Norvig, 2021). These aspects foster trust in AI systems by allowing their users to understand their behavior, limitations, and post-mortem assessment (Dignum, 2019). Addressing these aspects in the military cyber domain, this research aims at defining and analysing essential elements and methods relevant for the design and application of XAI models when targeting in military cyber operations. Hence, this research contributes to the existing body of knowledge in the AI, military, and cyber security domains, and calls for further research in this direction while also representing a basis for further design of strategies, policies, methods, and techniques applicable in the military cyber domain.

This research continues by analysing the challenges, opportunities, and effects of developing XAI models in the military cyber domain in a multidisciplinary setting with concrete cases and modelling solutions. Additionally, this research stresses the need for stakeholders' education and awareness with dedicated programs and courses, and the need of developing modelling and simulation, gaming, test beds, and digital twin solutions that facilitate and strengthen the responsibility, transparency, and fairness of i) the stakeholders involved in developing XAI models in the military cyber domain, and ii) the XAI models themselves.

References

Adadi, A. and Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE* Access.

Alonso, J. M. (2020). Teaching Explainable Artificial Intelligence to High School Students. *International Journal of Computational Intelligence Systems*, *13*(1), pp. 974-987.

Arrieta, A. B. et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, pp. 82-115.

Atakishiyev, S. et al. (2020). A multi-component framework for the analysis and design of explainable artificial intelligence. *arXiv preprint arXiv:2005.01908*.

Boothby, W., H. and Schmitt, M., N. (2012). *The law of targeting*. Oxford University Press.

Bistron, M. and Piotrowski, Z. (2021). Artificial Intelligence Applications in Military Systems and Their Influence on Sense of Security of Citizens. *Electronics*, *10*(7), 871.

Board, D. I. (2019). AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the *Department of Defense: Supporting Document.*

Deeks, A. (2019). The judicial demand for explainable artificial intelligence. *Columbia Law Review*, *119*(7), pp. 1829-1850.

Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, *70*, pp. 245-317.

Cambridge Dictionary (2021). https://dictionary.cambridge.org/dictionary/english/explanation

Core, M. G., Lane, H. C., Van Lent, M., Gomboc, D., Solomon, S. and Rosenberg, M. (2006). Building explainable artificial intelligence systems. In *AAAI* (pp. 1766-1773).)

DARPA. (2016). Explainable Artificial Intelligence. https://www.darpa.mil/program/explainable-artificial-intelligence

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way*. Springer Nature.

Doran, D., Schulz, S. and Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv preprint arXiv:1710.00794*.

Gerlings, J., Shollo, A. and Constantiou, I. (2020). Reviewing the Need for Explainable Artificial Intelligence (xAI). *arXiv preprint arXiv:2012.01007*.

Google (2020). AI Explainability Whitepaper. *Google*. European Conference of the Prognostics and Health Management Society.

Gunning, D. and Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, *40*(2), pp. 44-58.

Guo, W. (2020). Explainable artificial intelligence for 6G: Improving trust between human and machine. *IEEE Communications Magazine*, *58*(6), pp. 39-45.

Hamon, R., Junklewitz, H. and Sanchez, I. (2020). Robustness and explainability of artificial intelligence. *Publications Office of the European Union*.

Hansen, L. K. and Rieger, L. (2019). Interpretability in intelligent systems–a new concept?. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, pp. 41-49

Hepenstal, S. and McNeish, D. (2020). Explainable Artificial Intelligence: What Do You Need to Know?. In *International Conference on Human-Computer Interaction*, Springer, pp. 266-275.

Hoadley, D. S. and Lucas, N. J. (2018). Artificial intelligence and national security.

Holzinger, A., Biemann, C., Pattichis, C. S. and Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain?. *arXiv preprint arXiv:1712.09923*.

Islam, S. R., Eberle, W. and Ghafoor, S. K. (2020). Towards quantification of explainability in explainable artificial intelligence methods. In The Thirty-Third International Flairs Conference.

Keneni, B. M., Kaur, D., Al Bataineh, A., Devabhaktuni, V. K., Javaid, A. Y., Zaientz, J. D. and Marinier, R. P. (2019). Evolving rule-based explainable artificial intelligence for unmanned aerial vehicles. *IEEE Access*, *7*, pp. 17001-17016.

Kolbasin, V. (2018). *How to explain predictions of your network?* Globallogic.

Lai, Y. S., Neill, D., Płoskoń, M. and Ringer, F. (2020). Explainable machine learning of the underlying physics of high-energy particle collisions. *arXiv preprint arXiv:2012.06582.*

Lehto, M. (2016). Theoretical examination of the Cyber Warfare environment. Proceedings of the International Conference on Cyber Warfare and Security, pp. 223-230.

Maathuis, C., Pieters, W. and Van Den Berg, J. (2016). Cyber weapons: a profiling framework. In *International Conference on Cyber Conflict 2016 (CyCon US)*, IEEE, pp. 1-8.

Maathuis, C., Pieters, W. & Van Den Berg, J. (2018a). A computational ontology for cyber operations. In *Proceedings of the 17th European Conference on Cyber Warfare and Security (ECCWS),* pp. 278-88.

Maathuis, C., Pieters, W. and Van Den Berg, J. (2018b). Assessment methodology for collateral damage and military (dis)advantage in Cyber Operations. In *International Conference on Military Communications (MILCOM)*, IEEE, pp. 1-6.

Maathuis, C., Pieters, W. and Van den Berg, J. (2020). Decision support model for effects estimation and proportionality assessment for targeting in cyber operations. *Defence Technology*. https://doi.org/10.1016/j.dt.2020.04.007.

Maxwell, P. (2020). Artificial Intelligence is the future of warfare (just not in the way you think), https://mwi.usma.edu/artificial-intelligence-future-warfare-just-not-way-think/

Meske, C., Bunde, E., Schneider, J. and Gersch, M. (2021). Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management*, pp. 1-11.

Mohseni, S., Zarei, N. and Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *11*, pp. 1-45.

Molnar, C. (2020). *Interpretable machine learning*. Lulu. com.

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K. and Grossman, D. (2020). *Military applications of artificial intelligence: ethical concerns in an uncertain world*. RAND.

NATO (2016). *NATO Standard AJP-3.9 Allied Joint Doctrine for Joint Targeting*. NATO Standardization Office.

Oxford Dictionary (2021). https://www.oxfordlearnersdictionaries.com/definition/english/explanation

Preece, A., Harborne, D., Braines, D., Tomsett, R. and Chakraborty, S. (2018). Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184.*

Preece, A., Braines, D., Cerutti, F. and Pham, T. (2019). Explainable AI for intelligence augmentation in multi-domain operations. *arXiv preprint arXiv:1910.07563*.

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Penguin.

Russell, S. and Norvig, P. (2021). Artificial Intelligence: A Modern Approach, Global Edition 4th. *Foundations*, *19*, 23.

Samek, W., Wiegand, T. and Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*.

Samek, W. and Müller, K. R. (2019). Towards explainable artificial intelligence. In *Explainable AI: interpreting, explaining and visualizing deep learning*, Springer, pp. 5-22.

Shukla, B., Fan, I. S. and Jennions, I. (2020). Opportunities for Explainable Artificial Intelligence in Aerospace Predictive Maintenance. In *PHM Society European Conference*, 5(1), pp. 11-11.

Stahl, B. C. (2021). *Artificial Intelligence for a Better Future: An Ecosystem Perspective on the Ethics of AI and Emerging Digital Technologies*, Springer Nature, pp. 124.

Streich, J. et al. (2020). Can exascale computing and explainable artificial intelligence applied to plant biology deliver on the United Nations sustainable development goals? *Current opinion in biotechnology*, *61*, pp. 217-225.

Svenmarck, P., Luotsinen, L., Nilsson, M. and Schubert, J. (2018). Possibilities and challenges for artificial intelligence in military applications. In *Proceedings of the NATO Big Data and Artificial Intelligence for Military Decision Making Specialists' Meeting*, pp. 1-16.

Swartout, W. R. and Moore, J. D. (1993). Explanation in second generation expert systems. In *Second generation expert systems*, Springer, pp. 543-585.

United States Army Joint Publications 3-12. (2018). *Cyberspace Operations*.

Vilone, G. and Longo, L. (2020). Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*.

Walsh, M. et al. (2021). *Exploring the Feasibility and Utility of Machine Learning-Assisted Command and Control: Volume 2, Supporting Technical Analysis*. Rand.