M&S support to operationalization of NATO principles of responsible use of AI

Jan Hodicky
Volkan Kucuk
Orzuri Rique
7857 Blandy Road, Norfolk, Virginia
USA

Jan.hodicky@act.nato.int

ABSTRACT

Modelling and Simulation (M&S) can be treated as a strategic asset that can incorporate artificial intelligence (AI) and Big Data to nurture realistic and agile models for creating on demand simulation-based experimentation. M&S powered with AI enables to treat as "black box" those entities whose characteristics and models are cumbersome, and support experimentation of complex environments at strategic level. Alternatively, M&S can be leveraged to understand and operationalise the recently approved NATO principles of responsible Use (PRUs) of AI in defence and check through simulation-based experimentation whether the AI solution is compliant with these six principles. This paper presents an AI powered use case that studies the relationship between the quality of military capabilities, percentage of gross domestic product (GDP) spent in military budget and the power index score of each nation. The operationalization of "explainability", which seeks to tailor design of the trained system to the explanations required by different stakeholders, is demonstrated. To do so, the data gathered from JANES database is kept constant and different AI models are used to provide explanations. The results allow to gain purpose to fit end- user understanding of the hidden interrelations between military preparedness and budgets, inter-alia.

1.0 INTORDUCTION

NATO's Artificial Intelligence (AI) [1] defines the core elements required to provide a baseline for an AI Ready Alliance. It represents the first in a series of technology-specific strategies following the agreement by Allies of NATO's Coherent Implementation Strategy on Emerging and Disruptive Technologies (EDTs) [2]; the second one being the NATO's Data Exploitation Framework Policy [3]. NATO's Artificial Intelligence Strategy defines six NATO Principles of Responsible Use (PRUs) of AI in Defence.

NATO Strategic Commands (SCs) were tasked in July 2021 by the Military Committee to develop NATO Next Generation Modelling & Simulation (NexGen). NexGen will provide an enhanced M&S capability to support decision-making, strategic studies, wargaming, training and education, defence and operational planning, and capability development. NexGen will have the ability to federate Modelling and Simulation (M&S) with the national M&S capabilities. It is expected that during its lifecycle, NexGen will integrate AI and Big Data.

On the one hand, NATO treats M&S as a strategic domain that can incorporate AI and Big Data to nurture realistic and agile models and create on demand simulation-based experimentation. These advanced M&S can be leveraged to build trust in new capabilities and plans, identify gaps as well as support current and future training activities. On the other hand, M&S can be used to operationalise the PRUs and check through simulation-based experimentation whether the AI tool is compliant with these six principles.

1.1 AI in NATO

NATO defines artificial intelligence as "the ability of machines to perform tasks that typically require human intelligence – for example, recognising patterns, learning from experience, drawing conclusions, making predictions, or taking action – whether digitally or as the smart software behind autonomous physical systems" [1].

NATO has a clear distinction between AI and Autonomy in weapon systems. Distinctively to AI, autonomy is defined as "the ability of a system to respond to uncertain situations by independently composing and selecting among different courses of action in order to accomplish goals based on knowledge and a contextual understanding of the world, itself, and the situation" [4]. NATO Allied Command Transformation (NATO ACT) also differentiates between "autonomous" systems and "automated" systems [5]. According to NATO ACT, autonomous functioning refers to "the ability of a system, platform, or software to complete a task without human intervention, using behaviours resulting from the interaction of computer programming with the external environment. Task or functions executed by a platform, or distributed between a platform and other parts of the system, may be performed using a variety of behaviours, which may include reasoning and problem solving, adaptation to unexpected situations, self-direction, and learning. Which functions are autonomous, and to the extent to which human operators can direct, control, or cancel functions, is determined by the system design trade-offs, mission complexity, external operating environment conditions, or legal or policy constraints". On the contrary, although they require no human intervention, automated functions "operate using a fixed set of inputs, rules, and outputs. Their behaviour is deterministic and largely predictable". To avoid possible interoperability issues, Allies and NATO partners should have a clear agreement about the implications of each of the terms. For this paper document, the AI definition provided by the NATO's AI Strategy[1] is followed.

In a late 2020 survey of 250-defence technology leaders for allied forces of NATO, all indicated that they were considering AI solutions for their armed forces, while 49% had already tested AI in some aspect of defence (29% implementing, 9% operating and 4% optimizing) [6][7]. 59% of defence organizations also reported to have an AI strategy and more broadly, 60% had a digital transformation strategy, but only 42% reported to have a framework for deploying AI ethically and safely. They also ranked 1) intelligence, surveillance, and reconnaissance (ISR), 2) semiautonomous vehicle enablement, 3) autonomous vehicle enablement, and 4) cyberspace operations as the main defence operations to which AI can potentially add value.

A more detailed state-of-the-art study about the adoption of AI and autonomy by the 30 Nations of the Alliance was provided by NATO's Cooperative Cyber Defence Centre of Excellence 2021 report [8]. Through a number of anonymous non-classified interviews, NATO-employed experts outlined that NATO holds the capacity to play the role of "facilitator" for an Alliance-level approach to military AI [9]. NATO's "facilitator" role will encompass hosting discussions between member states on military AI, guide to Allies' thinking on military AI, and provide collaboration opportunities on the development of military AI. NATO's longstanding operational and technical groundwork provides a basis more for AI adoption rather than development. A noteworthy NATO lead framework for AI adoption comprises AI ontologies and taxonomies, interoperability initiatives, military data management initiatives, and federated accreditation models for M&S, verification and validation, operational and material standards for man-machine and data teaming, inter-alia.

In a recent NATO questionnaire on AI sharing standards done by consultation, command and control staff (C3S), 12 Allies provided feedback and signalled their openness and willingness to engage in the development of AI artefact sharing standards and best practices [10]. However, they also highlighted security concerns in terms of data handling, ownership, privacy and vulnerability.

1.2 M&S in NATO

Modelling and Simulation (M&S) is a discipline that develops and/or uses models, simulations and simulation systems within its lifecycle. The M&S lifecycle starts with design of the model followed by its execution. Model is a physical, mathematical or otherwise logical representation of a system, entity, phenomenon, or process. Simulation is the execution of a system model over time, and finally, simulation system is a combination of interacting elements or components organized to provide a representation of a system or of a part of the real world for an intended use [11].

M&S is well established in NATO, through the policy level document, NATO M&S Master Plan version 2 [12]. The second part of the document identifies application areas where M&S can be applied. Basically, it can be divided into two main categories. The first one, training and education, is the most wide-spread M&S application area as almost all NATO nations have their own Training and Simulation Centre. Conversely, the second category, the M&S to support analysis, is somehow still cornered and needs to prove its value in NATO. Operations support, capability development, mission rehearsal and procurement are the examples from the second category.

Since 2021 NATO is developing the M&S specific programme called NexGen, which is envisioned as a cloud based web-enabled, single digital environment that is 24/7 available. It will be based on modular approaches that allow for rapid development for ever changing requirements such as cybersecurity, threats, interoperability challenges, data, and geo specific or geo typical terrain requirements to meet the user needs.

1.3 AI and M&S gap analysis summary

A general assessment of the maturity analysis of the DOTMLPFI dimensions for AI-enabled M&S within NATO done in ACT is presented next. The most advanced dimensions are material and facilities, as Secure Data Science & AI exploration, experimentation and development (SANDI) and NATO Software Factory (NSF) environments will enable to have the infrastructure and software needed to developed AI-enabled M&S. In addition, in terms of personnel, the Allies have expressed their willingness to participate in AI collaborative computing environment, although some of them have limited resources and expertise. Most of the focus to increase the maturity should be put on doctrine, organization, training, leadership and education and particularly, in interoperability. Allies expressed their appreciation for standardization for the purpose of interoperability. However, there is a need to revise and adapt current M&S mandated standards such as STANAG 4603 and IEEE 1730 to account for AI and Big Data. To succeed in these efforts, the M&S and AI communities could further leverage events such as Coalition Warrior Interoperability eXercise (CWIX).

The limitations associated with AI are not only technological but also emerge through inter-sections with political, social and economic conditions. Real-world military data tends to be highly classified, and there may not be enough to adequately train an AI system. Instead, many military AI algorithms would need to be trained on simulation data, which may not accurately represent the real world, especially for safety-critical systems [8]. In addition, depending on the strategic, operational or tactical level use of M&S, the amount of available data for training purposes may be diverse. Consequently, a clear definition of “Big Data”, “Medium Data” and “Small Data” sets for different applications of M&S must be developed and its possible implications analysed.

There is also a necessity to develop models that are resilient to “data poising”. “Data poising” happens when an adversary introduces bad data to the algorithm so that it learns incorrect information. Even subtle changes to data can have big effects on an algorithm’s performance and in its subsequent use on simulations. Therefore, Validation, Verification & Accreditation (VV&A) procedures that check for data vulnerabilities for both, the trained models and the simulations in where they are used, should be developed.

In addition, “digital passports” [13] based on updated metadata standards need to be developed for cataloguing AI trained models and simulations that are used into. This initiative will help faster standardization and support

interoperability. It will also allow tracking whether the AI trained models and AI-enabled simulations have passed the NATO PRUs compliance certification or security certification requirements like threat analysis frameworks. The "digital passport" may also compile all the lifecycle historical data of the AI in case audits are requested. It is noted that the lifecycle of AI trained models and AI-enabled simulations comprises the design, development, operational and dismissal phases and in case of the simulations, the scope of each phase could be different depending on if the AI piece was introduced from its inception or a posteriori.

In terms of training and education, there is a necessity to educate and enhance the knowledge of AI algorithms and AI-enabled simulations. Humans often have "automation bias" whereby they over trust computer systems, even when they know they are flawed. Furthermore, it is cornerstone to define where it is possible to switch from "human in the loop" to "human on the loop" or "human over the loop", meaning that there is also a need to describe the human involvement in the different stages of the AI-enabled simulations. This transition will require some time so that "calibrated trust" is built. Commanders need to know how those systems have been developed and under what circumstances have been tested and used. To achieve this purpose, testing and evaluation needs to be integrated early in the development process up to they are operative.

2.0 M&S SUPPORT TO OPERATIONALIZE NATO PRUS

The six NATO PRUs of AI in Defence identified in the NATO's AI Strategy [1] are lawfulness, responsibility and accountability, explainability and traceability, reliability, governability and bias mitigation. The Allies in the AI Strategy agreed the definition of each of the PRUs, but they may not match one-to-one with national definitions or the ones available in the PMESII (Political, Military, Economic, Social, Infrastructure and Information) domain. NATO is currently working on the operationalization of these PRUs and has several initiatives, including a NATO Industrial Advisory Group (NIAG) study. The key difference between military AI and civilian AI is that in military AI you are not able to involve the targets. Therefore, critical questions of selection and engagement of targets needs to be answered including if the AI tool is doing better that the human baseline and where this baseline lies. For instance, human degrade the performance under stress and in case we want to improve that, it may be necessary to think about the development of AI models that reduce the collateral damage of the theatre of war. The training dataset for these AI models can be generated from simulation tools so that the AI trained model does not capture the poor human performance under stress. These simulation tools can be supported by military specific ontology for AI so that the check of NATO PRUs are integrated early in the process.

In addition, the systems are generally characterized by the average behaviour. Nonetheless, variability has also an important role to play as systems can be less variable than humans. Human performance is little known as there is not a good traceability of the human decision-making. Developing AI models that surface hidden patterns of the decision making process could improve the traceability of the process.

Adopting civilian systems can cause to propagate the bias into military systems. Bias could come from the dataset, from the AI trained model or from the final system itself and thus, there is a need to have a systematic approach to check for this bias. There are currently commercial open-source toolboxes [14][15], that have models that automatically check for possible bias of the AI system. These models can be modified and tested within the military domain so that the different layers of the AI enabled simulations (i.e. data, AI trained model, and AI enabled simulation itself) are compliant with NATO PRUs.

Furthermore, when it comes to understanding how AI enabled systems work, there could be a frustration due to the "blackbox" approach, and specifically when it comes to the experimentation, testing and evaluation phase of an assisted decision making system. The "explainability" PRU seeks to tailor the outcome of the trained system to the explanations required by different stakeholders, as different stakeholders require explanations for different purposes and with different objectives. The use case in this paper presents a way to describe the operationalisation of explainability for assisted decision making at different level of command.

In summary, M&S can be leveraged to add models on top of the original model not only to gain insight in how the AI system work but also to create NATO PRUs compliant tools. Simulation can also be applied to create PRU compliant data when other means to gather these data are not available or to test the operationalization of PRUs with different users and under different scenarios. Each of the NATO PRUs may have different significance in different applications and hence, different weights may be needed for each M&S. M&S can support the quantification of these weights and as feasibly can alter through models the weight of each PRU and gather the feedback of different stakeholders when these models are applied in AI-enabled simulations.

3.0 OPERATIONALIZATION OF EXPLAINABILITY USE CASE

The AI powered use case studies the relationship between the quality of military capabilities, percentage of gross domestic product (GDP) spent in military budget and the power index score of each nation. The operationalization of "explainability", which seeks to tailor the design of the AI model and the aggregation of input data to the explanations required by different stakeholders, is demonstrated. To do so, the data gathered from JANES [17] and Global Firepower Ranking [16] are used as the unique data set to different AI models providing the purpose to fit end- user understanding of the hidden interrelations between military preparedness and budgets, inter-alia.

Therefore, our main presumption for explainability is that different stakeholders require explanations for different purposes and with different objectives, and explanations needs to be tailored to their needs.

There are three levels of the end-user working with the proposed prototype system. The highest level is represented by political will, it correspond to NATO Atlantic Council (NAC), the second is military strategic level corresponding to ACT or ACO and the lowest considered it the military operational level.

3.1 JANES Database

JANES is a subscription-based service that provides nations' military capabilities and basic financial indicators that are collected from open sources. The site also provides API access to easily collect data, but in our case this option was not used.

3.2 Global Firepower Ranking

Global Firepower Ranking is a web site that publishes nation's overall military power scores (defined as *PowerIndex* rest of this paper) and rankings annually. Basically, the nations' ranking are based on order of *PowerIndex* values of nations. The algorithms and models that Global Firepower uses to calculate *PowerIndex value* are not publicly available, therefore in their website it is stated as following;

"The finalized Global Firepower ranking below utilizes over 50 individual factors to determine a given nation's PowerIndex ('PwrIndx') score with categories ranging from military might and financials to logistical capability and geography."

As a result, the calculation methods and the "50 individual factors" are "black boxes" for researchers, but calculated PowerIndex value (PIV) is publicly available. It is also stated in their website that, 0 (zero) value for the PIV is a perfect score, which means that the more closer to 0 (zero), the more power a country has. (i.e. the USA has a PIV score 0.00453, and 1st Ranking).

3.3 Creation of Dataset and Pre-Processing

In order to create a dataset for our research, we have collected 21 types of military capabilities and 2 types of financial indicators (Table 3-1) of 31 nations (*17 NATO nations, 14 Non-NATO Nations*) from JANES database. In JANES, the individual military capabilities of each nation are represented by color codes, which

are black, yellow, orange and green, respectively. (*Meanings: Black- There is no capability to Green-The capability is sufficient*). Color codes were converted in to numeric scale (0-black, 1-red, 2-orange, 3-yellow, 4-green) and the data was normalized in to 0-1 scale for the sake of simplicity. Furthermore, the two financial indicators total defence budget (TDB) and percentage of gross domestic product (GDP) are also collected and normalized.

Additionally, PowerIndex (*PIV*) score of the 31 nations from Global Firepower Ranking website was also collected and normalized, and combined JANES and Global Firepower Ranking values into single dataset. (*Full dataset structure is shown in Table 3-1*)

In order to define a threshold value to classify nations, we selected the average value of *PIV* score of all NATO Nations in our dataset, which is *0.2659. T*his value is used as a reference for labelling (***Fit4Purpose***) for the all the nations in our dataset. If a nation's *PIV* score is below threshold, then the nation is labelled as '1', otherwise '0. This value is used for further analysis in upcoming sections.

$$Fit4Purpose_{Nation_j} = \begin{cases} 1, & if\ Nation_j(PIV) < 0.2659 \\ 0, & otherwise \end{cases}$$

After creation, the dataset was divided into training and testing subsets, in which 20 nations (*11 NATO Nations*) were selected for training and 11 nations (*6 NATO Nations*) for testing, respectively.

Table 3- 1 Input Data Types for Proposed Model

MCi	Military Capability (*MC*)	Capability Domain
1	Air-to-Air Warfare	Air Defence
2	Ground Based Air Defence	Air Defence
3	Maritime Anti-Air Warfare	Air Defence
4	Offensive Air Support	Fire Support
5	Indirect Fire	Fire Support
6	Naval Surface Fire Support	Fire Support
7	Air (Space) Recon	ISR
8	Ground Recon	ISR
9	Maritime Surveillance	ISR
10	Aerial Refuelling	Logistics
11	Airlift	Logistics
12	Maritime Transport	Logistics
13	A SuW – Airborne	Anti-surface Warfare
14	A SuW – Surface	Anti-surface Warfare
15	A SuW- Submarine	Anti-surface Warfare
16	ASW – Airborne	Anti-Submarine Warfare
17	ASW – Surface	Anti-Submarine Warfare
18	ASW- Submarine	Anti-Submarine Warfare
19	Armoured Warfare	Direct Ground Combat
20	Infantry Ops	Direct Ground Combat
21	Combat Engineering	Direct Ground Combat

22	Total Defence Budget - TDB	Financial
23	Gross Domestic Product % – % GDP	Financial
24	Global Firepower Ranking	PIV Score from website – Used for Regression
25	Fit4Purpose	Label - Generated for Classification (1/0) purposes based on threshold value (0.2659).

3.2 AI Applications on Dataset

To demonstrate explainability, 4 different AI models on our dataset with different levels of details were created as follows:

1) Neural Network with All Features for Baseline comparison
2) Explainability by Linear Regression (LR) - All Features in Dataset for Operational Level User
3) Explainability by Linear Regression (LR) - Domain-Based Features for Strategic Level User
4) Explainability by Linear Regression (LR) - Fully Aggregated Military Capabilities and %GDP

WEKA [18] was used to create AI models conduct tests. WEKA is an open source, Java based software that provides a collection of machine learning algorithms for data mining tasks.

3.2.1 Neural Network with All Features for Baseline comparison

At first setup, we applied Neural Network (NN) to our whole dataset in order to classify nations. For setup, we used all the features except *Global Firepower Ranking* (*MC_{24} in Table 3-1*) as inputs, and, Fit4Purpose label as output, respectively, see Figure 3-1

By creating a NN structure with 4 hidden layers, 0.3 learning rate, 0.2 momentum and 1500 epochs, we managed to classify 10 nations out of 11 correctly. (90.90% of accuracy), see Table 3-3 and the grey cell. The detailed accuracy rates of proposed NN setup are presented in Table 3-2.

Table 3- 2 Detailed accuracy Rates of selected NN

(4 Hidden Layers, Momentum: 0.2, Learning Rate: 0.3, Epochs: 1500)

TP	FP	Precision	Recall	F-Measure	Class
0.750	0	1	0.75	0.857	0
1	0.25	0.875	1	0.933	1

Table 3- 3 Predictions of selected NN

Country	*PIV*	Class based on *PIV*	Predicted Class (NN Result)	Country	*PIV*	Class based on *PIV*	Predicted Class (NN Result)
1	1	0	0	7	0.002	1	1
2	0.032	1	1	8	0.002	1	1
3	0.072	1	1	9	0.145	1	1
4	0.052	1	1	10	0.329	0	1
5	0.855	0	0	11	0.108	1	1
6	0.638	0	0				

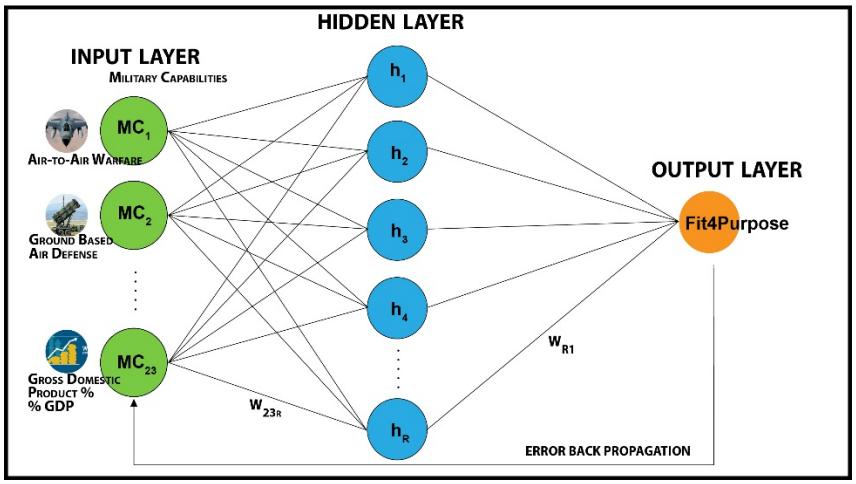

Figure 3-1 NN Representation

By Baseline model we are not getting further details related to explainability. Therefore following models were implemented with special focus on their operational level needs.

3.2.2 Explainability by LR - All Features in Dataset for Operational Level User

An operational level user is interested in analysing the data as much in detail as possible. To present a detailed view for that kind of users, we used whole dataset shown in Table-3.1 (except Fit4Purpose value) to train our model by Linear Regression. The LR model was trained based on *PIV* values in our training dataset. We defined the calculated value of a test instance as *CPIV (Calculated Power Index Value)* as

$$CPIV = a_0 + w_1.MC_1 + \cdots w_{23}.MC_{23},$$

where *MC* refers to Military capability value, and *w* refers to weight of *MC.*

The weights of our LR model for Operational Level model are shown in Table-3.4. These weights should create the main foundation for the explainability.

Table 3- 4 Linear Regression Model Weights for Operational Level Model

i	Capability (MC_i)	Weight (w_i)	#	Capability (X_i)	Weight (w_i)
1	Air-to-Air Warfare	-0.14704	13	A SuW – Airborne	0.13195
2	Ground Based Air Defence	-0.11267	14	A SuW – Surface	-0.02302
3	Maritime Anti-Air Warfare	0.46032	15	A SuW- Submarine	-0.20917
4	Offensive Air Support	-0.22945	16	ASW – Airborne	0.16949
5	Indirect Fire	-0.29129	17	ASW – Surface	-0.42871
6	Naval Surface Fire Support	0.02939	18	ASW- Submarine	-0.03204
7	Air (Space) Recon	0.08968	19	Armoured Warfare	0.28709
8	Ground Recon	0.14598	20	Infantry Ops	0.19512
9	Maritime Surveillance	0.15639	21	Combat Engineering	-0.17213
10	Aerial Refuelling	-0.11916	22	Total Defence Budget - TDB	-0.03878
11	Airlift	0.08507	23	Gross Domestic Product – GDP	-0.00086
12	Maritime Transport	-0.27631		Intercept (α)	0.32289

Table 3- 5 Results of Linear Regression Model for Operational Level Model

Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root relative squared error	Accuracy based on *CPIV* Classification (threshold: 0.2659
0.8218	0.1981	0.2523	75.65%	67.32%	100%

Table 3- 6 Predictions and Comparison with NN Model

Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)	Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)
1	0.363	0	0	7	-0.095	1	1
2	0.123	1	1	8	0.215	1	1
3	0.112	1	1	9	0.006	1	1
4	-0.047	1	1	10	0.213	1	1
5	0.585	0	0	11	-0.113	1	1
6	0.381	0	0				

Table 3-5 and Table 3-6 describe the comparison between the baseline NN model and the current LL model. In this case the match between these two models were 100%.

3.2.4 Explainability by LR - Domain-Based Features for Strategic Level User

To reduce the complexity of input data, we combined military capabilities of our dataset into domain level. Rather than dealing with 21 military capability features, we combined these capabilities in to 7 domain features by calculating the average value of each domain. For instance, Air Defence Domain value is average of Air-to-Air Warfare, Ground Based Air Defence and Maritime Anti-Air Warfare values (see Table3-7). Two financial indicators (TDB, % GDP) are included individually. Again, the LR model was trained based on *PIV* values in our training dataset. We defined the calculated value of a test instance as *CPIV* as

$$CPIV = a_0 + w_1.CD_1 + \cdots w_9.CD_9$$

where *CD* refers to *Capability Domain* value, and *w* refers to weight of *CD*. The weights of our LR model for Operational Level model are shown in Table-3-8.

Table 3- 7 Inputs for Strategic Level User

CD_i	Capability Domain (*CD*)	Included Military Capability
1	Air Defence	Air-to-Air Warfare
		Ground Based Air Defence
		Maritime Anti-Air Warfare
2	Fire Support	Offensive Air Support
		Indirect Fire
		Naval Surface Fire Support
3	ISR	Air (Space) Recon
		Ground Recon
		Maritime Surveillance
4	Logistics	Aerial Refuelling
		Airlift
		Maritime Transport
5	Anti-surface Warfare	A SuW – Airborne
		A SuW – Surface
		A SuW- Submarine
6	Anti-Submarine Warfare	ASW – Airborne
		ASW – Surface
		ASW- Submarine
7	Direct Ground Combat	Armoured Warfare
		Infantry Ops
		Combat Engineering
8	Total Defence Budget - TDB	-
9	Gross Domestic Product % – % GDP	-

Table 3- 8 Linear Regression Weights for Operational Level Model

i	Capability (MC_i)	Weight (w_i)
1	Air Defence	-0.14704
2	Fire Support	-0.11267
3	ISR	0.46032
4	Logistics	0.02939
5	Anti-surface Warfare	0.08968
6	Anti-Submarine Warfare	0.08507
7	Direct Ground Combat	-0.27631
8	Total Defence Budget - TDB	-0.1416
9	Gross Domestic Product % – % GDP	-0.1037
	Intercept (α)	0.4136

Table 3- 9 Linear Regression Weights for Operational Level Model

Table 3- 10 Predictions and Comparison with NN Model

Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)	Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)
1	0.362	0	0	7	0.001	1	1
2	0.159	1	1	8	-0.003	1	1
3	0.232	1	1	9	0.160	1	1
4	0.164	1	1	10	0.313	0	1
5	0.419	0	0	11	0.193	1	1
6	0.340	0	0				

Table 3- 11 Results of Linear Regression Model for Strategic Level Model

Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root relative squared error	Accuracy based on *CPIV* Classification (threshold: 0.2659
0.8469	0.1726	0.2607	65.92%	69.55%	90.90%

Table 3-9 and Table 3-10 describe the comparison between the baseline NN model and the current LL model. In this case the match between these two models were 90.90%.

3.2.5 Explainability by LR - Fully Aggregated Data Set for Political Level User

For the political level, our aim was to reduce the complexity. For this purpose, we aggregated all the military capabilities in to one single military capability value, which was determined by getting the average of all military capabilities. Furthermore, to differentiate financial indicators, we used only % GDP value in political level model, see Table 3-11. LR model was trained based on *PIV* values in our training dataset. We defined the calculated value of a test instance as

$$CPIV = a_0 + w_{CMP}.CD_{CMP} + w_{GDP}.CD_{GDP}$$

where *CMP* refers to *Combined Military Power* value, and *w* refers to weight of *CMP*. The weights of our LR model for Political Level model are shown in Table-3-12.

Table 3- 12 Inputs for Political Level User

#	Input	Information
1	Combined Military Power (*CMP*)	Average of all military capabilities
2	% Gross Domestic Product % – % GDP	-

3.3.4.1 Weights of LR Model

Table 3- 13 Linear Regression Weights for Political Level Model

#	Input	Weight (w_i)
1	CMP	-0.534
2	% GDP	-0.1713
	Intercept (α)	0.4829

3.3.4.2 Predicted Values and Comparison with NN Model (Section 3.2.1)

Table 3- 13 Predictions and Comparison with NN Model

Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)	Country	*CPIV* – LR Result	Class based on *CPIV* (threshold: 0.2659)	Predicted Class (NN Result)
1	0.447	0	0	7	0.099	1	1
2	0.1	1	1	8	0.133	1	1
3	0.109	1	1	9	0.151	1	1
4	0.101	1	1	10	0.22	1	1
5	0.422	0	0	11	0.392	0	1
6	0.421	0	0				

3.3.4.3 Results of LR Model

Table 3- 14 Results of Linear Regression Model for Political Level Model

Correlation Coefficient	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Root relative squared error	Accuracy based on *CPIV* Classification (threshold: 0.2659
0.8488	0.1771	0.2537	67.64 %	67.71 %	90.90 %

Table 3-13 and Table 3-14 describe the comparison between the baseline NN model and the current LL model. In this case the match between these two models were 90.90%.

4.0 CONSLUSION

The use case describes a proposed way to support explainability based on the implementation of a different AI models designed based on the specific user needs but founded in the same data set. We have created the baseline model using the NN approach to describe the relation between military capabilities and national budget. Then trying to bring arguments for explanability, we have implemented three different AI models. These three AI models used the same data set, but we aggregated the inputs from the data set for their training. We reduced the complexity of the models based on the level of details that particular user can be looking for. Therefore the model for the political level has the lowest complexity and brings relatively simple explainability and vice versa.

In our use case, the similarity between AI models classification predictions was pretty high. We reached over 90.90% match. If a match between models outputs is sufficiently high then models should be mutually used to support their behaviour explanation.

The prototype developed based on the AI models can serve different purposes, e.g. to evaluate the power of the country that is applying to enter NATO. The value of the threshold in the classification algorithm, is driving the type of question for the end-user what if analysis.

There are limits and constrains in our use case. The amount of the data used for training period is critical. However, it does not degrade the general recommendation for the AI models developers.

By this use case, for the sake of explainability, we are proposing that developers should be implementing more than just one AI model of the problem domain. They should be actively looking for implementing other AI models that would bring similar results from the same dataset but creating the value for explainability.

REFERENCES

[1] PO(2021)0350, NATO's Artificial Intelligence Strategy (OCT 2021)

[2] PO(2021)0059, Foster and Protect: NATO's Coherent Implementation Strategy on Emerging and Disruptive Technologies (FEB 2021)

[3] PO(2021)0360, NATO's Data Exploitation Framework Policy (OCT 2021)

[4] NATO Science & Technology Organization, "Science & Technology Trends 2020-2040. Exploring the S&T Edge". March 2020

[5] A. Kuptel and A. Williams, 'Multinational Capability Development Campaign (MCDC) 2013-2014, Focus Area 'Role of Autonomous Systems in Gaining Operational Access,' Policy Guidance: Autonomy in Defence Systems', NATO Allied Command Transformation, 29 October 2014

[6] D. Chenok, L. van Bochoven and D. Zaharchuk, "Deploying AI in Defense Organizations", IBM Center for the Business of Government, 2021, URL: https://www.ibm.com/downloads/cas/EJBREOMX (Last accessed 2021/12/13)

[7] K. Chandler, "Does Military AI Have Gender? Understanding bias and promoting ethical approaches in military applications of AI", UNIDIR, Geneva 2021. URL: https://doi.org/10.37559/GEN/2021/04 (Last accessed 2021/12/13)

[8] M. Gray, A. Ertan, "Artificial Intelligence and Autonomy in the Military: An Overview of NATO Member States Strategies and Deployment", Tallinn 2021, URL: https://ccdcoe.org/uploads/2021/12/Strategies_and_Deployment_A4.pdf (Last accessed 2022/01/24)

[9] Z. Stanley-Lockman, 'Military AI Cooperation Toolbox', August 2021, URL: https://cset.georgetown.edu/publication/military-ai-cooperation-toolbox/ (Last Accessed 2022/01/24)

[10] E. Grissom, "Initial Findings of NATO Questionnaire on AI Sharing Standards", TIDE Sprint presentation, Sopot (Poland), April 2022

[11] AAP-06. Edition 2021. NATO glossary of terms and definitions

[12] NATO ACT. NATO Modelling and Simulation Master Plan. Edition 2. 2021. AC/323/NMSG(2012)-015

[13] C. Gorski, "Informal Virtual Workshop: Operationalising NATO Principles of Responsible Use for AI in Defence", March 3, 2022

[14] IBM, AI Fairness 360, URL: http://aif360.mybluemix.net/ (Last accessed 2021/12/17)

[15] Microsoft, Responsible AI resources, URL https://www.microsoft.com/en-us/ai/responsible-ai-resources (Last accessed 2022/02/24)

[16] Global Firepower Homepage, URL https://www.globalfirepower.com/, (Last accessed 2022/08/02)

[17] JANES website, URL: https://www.janes.com/ (Last accessed 2022/08/08)

[18] WEKA website, URL https://www.cs.waikato.ac.nz/ml/weka/ (Last accessed 2022/08/08)