Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 55-29 (2022) 25–30

Towards Generation of Synthetic Data Sets for Hybrid Conflict Modelling

Sven Nõmm [*,**] **Adrian Venables** [*,**]

[*] *SensusQ, Tellliskivi str 60, Tallinn, Estonia, {sven, adrian }@ sensusq.com)*

[**] *Department of Software Science, School of Information Technologies, Tallinn University of Technology, Akadeemia tee 15a, 12618, Tallinn, Estonia (e-mail: {sven.nomm, adrian.venables }@ ttu.ee).*

Abstract: Design proposal for an AI-driven military situational awareness application. Current events in Ukraine have emphasized the importance of the Information Environment in supporting military operations. Activities in the physical domain are processed in the virtual domain of computers and networks before being interpreted by the human cognitive domain where decisions are made. Commanders at all levels have an ever increasing amount of information of varying latency and reliability available to them from a variety of sources. These range from highly sophisticated and complex bespoke surveillance systems to individuals equipped with a smartphone and Internet connection. An effective commander must be able to assimilate all the information sources available to them and be able to visualise the battlespace in an accurate and timely manner. To assist them, there are already a wide variety of software applications capable of receiving multiple inputs and displaying the disposition of military forces within a mapping environment. However, these tools place the interpretation and analysis responsibilities with the human operator. This paper proposes an AI driven process in which the initial analysis and correlation function is conducted in real time and in response to data inputs from multiple sources. This presents the decision maker with a fused, correlated and predictive Common Operational Picture providing a clear information advantage.

Keywords: Hybrid conflict, Machine learning, synthetic data

1. INTRODUCTION

This paper focuses on the challenge of data set generation to benchmark and validate Artificial Intelligence (AI) based techniques for the detection and correlation of events in social media. The use of the Information Environment to influence target audiences is a component of every current conflict as part of a wider hybrid warfare strategy. In spite of there being no commonly accepted definition of hybrid conflict, different definitions have been proposed Thiele (2015), Cayirci et al. (2016) and numerous modelling and simulations attempts have been made Lieberman (2012). Some of these proposals have led to the requirement to model behavioural and social aspects of the target population or social group Sokolowski and Banks (2007), Sokolowski and Banks (2009). The complexity of systems designed to model hybrid conflicts was discussed in Kott and Corpac (2007). One of these possible paradigms is described in Balaban and Mielniczek (2018). In response to the increasing range of information sources, the development of applications to interrogate the properties of the Information Environment in non-military and non-political conflicts is now becoming recognised. The dynamics and methods of these operations are very similar to those used in traditional kinetic conflict. This is especially true regarding the usage of social and mainstream media (MSM) channels. For example, social media may be actively used not only to influence public opinion on political or military issues but also to encourage actions such as brand boycotting or inciting panic for money withdrawal from ATMs. Malicious actors in the Information Environment can create an enormous amount of data for analysts to process. Friendly and neutral users document events as they occur, with adversaries possibly employing troll farms and AI-controlled bots in an attempt to organise and control the content of media organisations. Attackers are also using AI to control bot networks, whilst defenders can utilise AI's attributes to detect attacks in their preliminary stages and prioritise countermeasures. Therefore, AI techniques have an increasingly important role in presenting a timely, accurate and comprehensive common operational picture. An AI-based approach to data interpretation is becoming an increasing trend in many areas of cyber security and related disciplines. For example, statistical machine learning techniques Bahşi et al. (2018) and deep learning methods Meidan et al. (2018) have demonstrated highly accurate results in detecting and classifying botnet attacks. However, such techniques require large data sets for training and validation. While hybrid operations are not a new phenomenon, available data sets describing such operations may not necessarily meet the requirements of training and validating general-purpose attack detectors. As hybrid operations are tailored for each particular use case, highly biased and skewed data sets may cause significant issues. This can be resolved by the generation of synthetic data sets that allow complete user control over

10.1016/j.ifacol.2022.10.226

any data property and so may be seen as the answer to the absence of real world data. Generating accurate and useful synthetic data sets requires an intimate understanding of actual hybrid operations and, wherever possible, uncover the mathematical nature of underlying processes. For precise targeting of the present contribution, the authors refer to Unknown group of authors (2018), which requires conducting a simulation (war game). Furthermore, if the AI component is involved, one is obliged to provide large data sets for training and validation purposes, which constitute the subject of the present research. This paper is organised as follows. Section 2 explains the necessary background of hybrid operations and formalises the motivation of the present research. Section 3 provides a breakdown of the operations into actions, events and components and, whenever possible, describes the behaviour of the corresponding processes. The structure of the proposed generator is discussed in section 4. A proposed solution and its limitations are discussed in section 5. Concluding remarks are stated in the last section.

2. MOTIVATION AND BACKGROUND

Currently, the analysis of hybrid and influence operations and their MSM and social media components are mostly performed by human analysts. However, the increasing number of available data sources combined with the increasing employment of bots and AI-based techniques have led to large volumes of information to process. In turn, this makes it harder for a human to perceive, filter and use the information to make adequate timely decisions. Also, the involvement of adversary AI based systems increases the necessity for faster response or action times. This, in turn, further increases the overall pressure on the analysts. A logical decision to overcome the limitations of human operators is to provide AI-based tools to filter and process the information gathered by observing media and social media channels. However, AI tools require large data sets for training, validation and bench marking. It is important to note that available data sets describing a particular conflict or operation may be biased and not necessarily sufficiently diverse for the training and validation of AI models. This leads to the idea of synthetic data generator allowing appropriate artificial data to be produced that can compensate for the lack of suitable real world data.

3. STRUCTURAL ANALYSIS

In considering the situation from the attacker's perspective, to perform any operation there are finite resources available. It is the role of the attacker's headquarters staffs to allocate these resources across a range of disciplines to achieve the optimal result. These can include operational security, manoeuvre, media operations, community outreach and engagement, psychological operations, deception plans, and kinetic fires when appropriate. These resource allocations may be altered during the course of the operation as priorities change. As the proportions change within the overall quantity of resources available, this will result in real world events and a media reaction. For example, Figure 1 illustrates how different actions or events will vary over a certain period, the entire operation length or for a particular phase and how they may be depicted. In attempting to view the entirety of a conflict, it may be that the defenders may not have sight of all events within the domains of operation. Also, observed information may be incomplete and alternative outcomes based on particular events may also be reported by different sources and be spread over time. For example, individual users may make social media posts independently and unrelated to the attacker or defender. There is also the phenomenon of *re-posting* which may or may not distort the importance of the information or cause some of its original meaning to be lost. Figure 2 depicts these events and the corresponding media occurrences view of the defending side. Consider now the events, their structure and relation to the environment. It is important to note here the potential for terminological confusion that may be caused by the usage of the word *event*. A probability theory event is defined as the set of the outcomes of an experiment. In the common 4 stage approach to decision-making known as the Observe, Orient, Decide, Act (OODA) loop, a event is an occurrence which is initially observed. This is filtered and put into context (Orient) before an appropriate Decision is made prior to Acting. This creates another event, which is an outcome with the attributes of time, location, means, organisations involved and actors involved. Further data

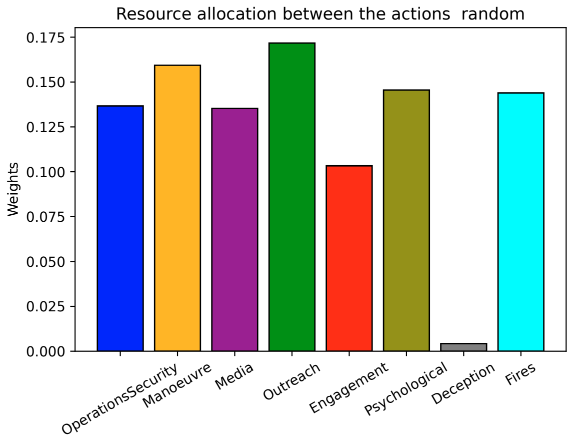

Fig. 1. Possible resource allocation between different actions

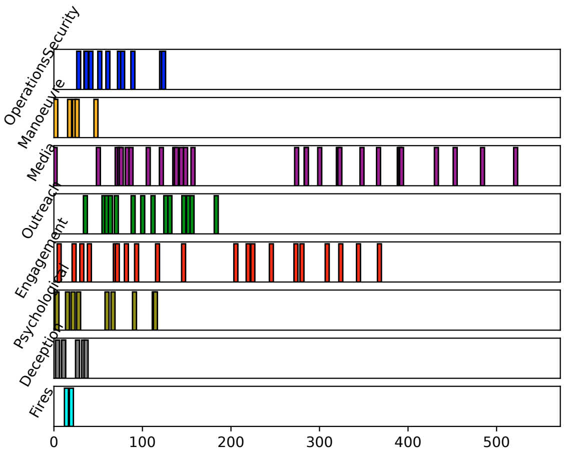

Fig. 2. Distribution of events along different actions.

Table 1. Example of an incomplete event information reported by different actors over a certain time interval.

Outcomes	Channel 1	...	Channel $n-1$	Channel n	Actual event
Event (what)	network disruption	...	Internet cable cut	Electric blackout	Bot net attack
Time (when)	few hrs. ago	...	now	hour ago	22:51 xx.xx.xxxx
Location (where)		...	Corner of boulevard and 5th str.	5th str.	Internet provider
Actors		...	Tractor driver	Electric supplier	Unknown hackers
Organisations		...	Repair subcontractor		Unknown
Means		...	Tractor		No information

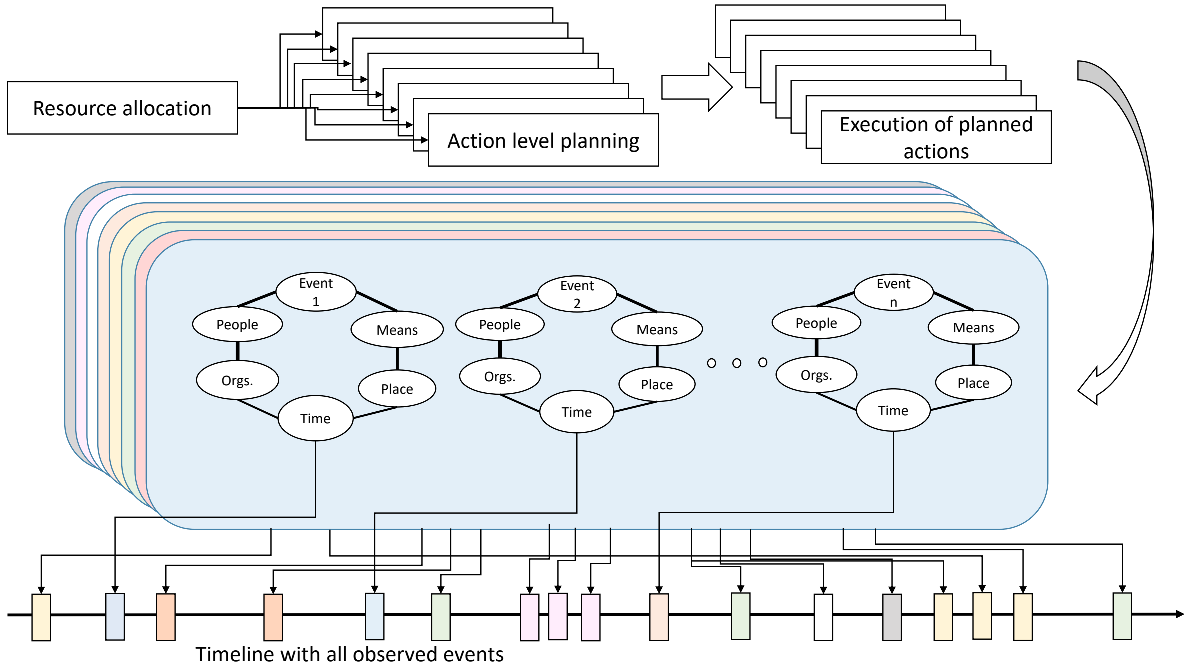

Fig. 3. From action planning to events on the time line.

sources can be derived from military science that Hudson (1998), Unknown group of authors (2018) analyses and provides attributions to events using the system of *who*, *what*, *when*, *where*, *why* and *how* system of questions. In both cases care should be taken when referring to the precise meaning of the term *event*. Figure 3 illustrates the process of event generation from the attacking side viewpoint.

Let us suppose that event A has occurred. Let us further assume that the defending side is not aware of its occurrence. However, a few social media posts contain partial information about the event. Table 1 depicts a simplified example of an incomplete reporting of the same event. In a more formal setting, one may think about the event as the vector in the feature space in which elements are the experiment's outcomes. In an OODA setting, these are the values of time, place, actors, organisations and means, and in the military setting are the answers to the questions *who*, *what*, *when*, *where*, *why* and *how*. Social media posts may be seen then as the projection of the event into a different subspace of the information space containing the original feature set or even mappings to other feature spaces. However, it should be noted that social media users may not necessarily describe events correctly or describe them using their full vision and understanding. Figure 4 represent the view of the defending side.

There are two more issues that add complexity to the actions of defenders. The first is the re-posting of events. One or many posts describing the event may be re-posted completely, partially or even with some misinterpretation. The second issue is that algorithms in social media platforms influence the appearance of the posts in users' stream of content. Moreover, these algorithms are neither disclosed or are transparent for platform users or content creators Huszár et al. (2022). Topic modelling Skaza (2015)

4. PROPOSED DATA GENERATOR

The workflow implemented to generate the desired data set mimics the actual procedure for planning and executing hybrid operations. Here and after many planning steps, refer to the rules required to perform the particular step. Such rules are expected to be set by the human operator depending on the operation (mission) criteria, doctrine, training and experience. The usual practice is to answer the seven questions of the military command estimate Infantry Training Screw (2019). These are:

(1) What are the enemy doing and why?
(2) What have I been told to do and why?
(3) What actions/effects do I want to have on the enemy?
(4) Where can I best accomplish each action/effect?
(5) What resources do I need to accomplish each action/effect?

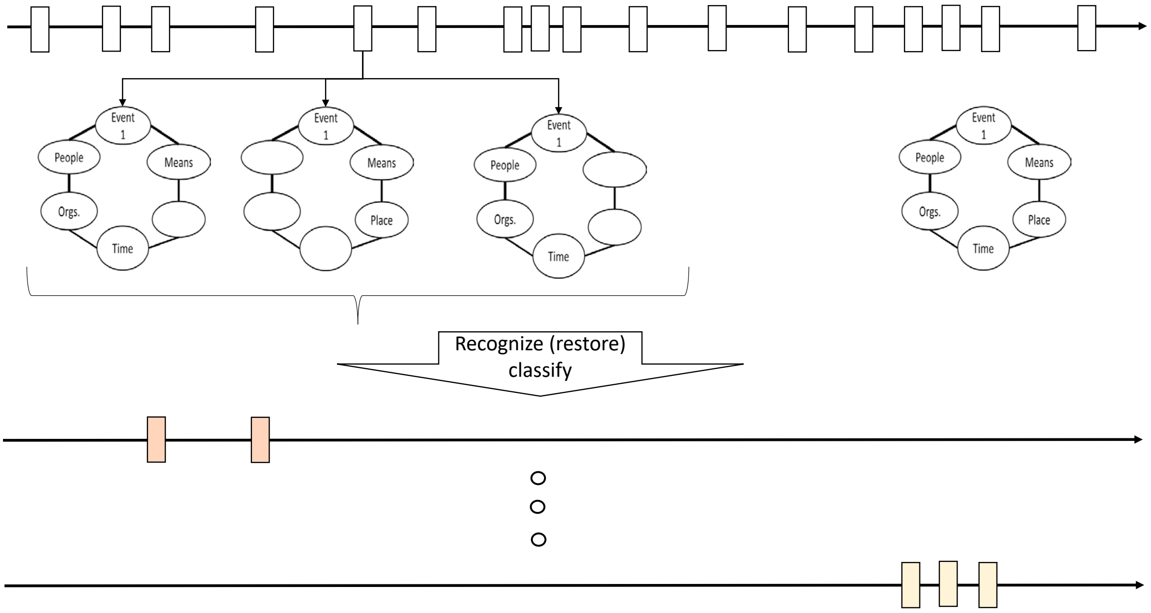

Fig. 4. Recognition and classification of the events

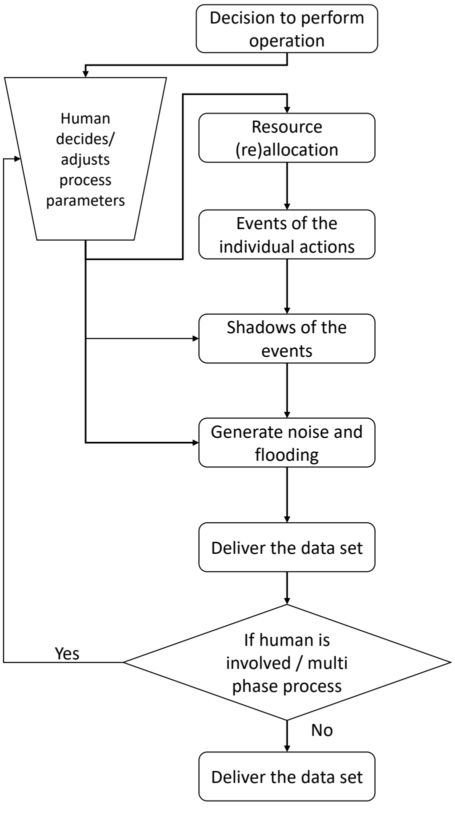
Decision to perform operation
Human decides/ adjusts process parameters
Resource (re)allocation
Events of the individual actions
Shadows of the events
Generate noise and flooding
Deliver the data set
If human is involved / multi phase process
Yes
No
Deliver the data set

Fig. 5. Flow chart of the proposed generator.

(6) When and where do these actions take place in relation to each other?
(7) What control measures do I need to impose?

Answers to these questions, matching actions, reports describing these actions, and inevitable copied and noise information can be generated by the workflow depicted in Figure 5.

(1) The resource allocation step does not differ much from the actual operations, with one exception. The resource allocated to the particular action is described by its weight (proportion) from the total resource allocated to the operation. This adjustment is made to simplify the mathematical representations used in the implementation. The human operator decides the particular rules of resource allocation.
(2) For each individual action, events and their location on the time line should be decided. In this case the rules of actual operations may be used. The rules here may be formulated in the form of instructions describing how many events of the given action should be carried out and how they should be distributed in time. At the same time from the perspective of the defending side these are the sequences of events which are modelled by probabilistic distributions. In the present version of the generator, the human operator may make a choice of a number of distributions or apply a rule based approach. Besides the probabilistic view of the event, one must complement it with outcomes inherited from the OODA paradigm or military science. Namely, one has to populate other outcomes: participants, means, organisations and place. To solve this part of the process, the generator relies first on the rule based procedure setting the restrictions regarding appearances of people, organisations and places with respect to the same event. To populate the remaining outcomes, the choice of the associative patterns generators is used Cooper and Zito (2007) or Omari et al. (2008). The latest are chosen because the means closely resemble market basket problems, and there are plenty of open-source generators to use. After this step all the events are positioned on the time line.

(3) Complete, incomplete and distorted re-posting of the news on different social media platforms creates many copies of the same news, referred to as *shadow of the event*. Shadow generation requires arguments and the first is the number of shadows with the second describing how long shadows will be generated. The current version of the generator uses *forgetting function* to model the length of time interval and inherits information about the resource allocation to the parent action to derive the number of shadows. Finally, different random functions are used to control the completeness and level of distortion of the shadow re postings. For the purposes of training the AI algorithm, each event is labelled. The labels allow identification of whether the event is real or is just a re-post of the original in social media.
(4) To generate the noise and flooding posts and re-posts, the operator should use the same generators but feed with events, actors, organisations and places from different databases.
(5) Finally, all generation information is merged into one dataset, with the events ordered in time.
(6) At this point, the data set may be delivered, or the process may be looped.

Figure 5 formalises the concept described above. Each action event and its location on the timeline should be decided. Each record of the returned dataset contains two groups of fields. The first group consists of the fields constituting actual information reporting.

`Event ID.`
`Timestamp.`
`Event name (description).`
`Reference to the place or coordinates.`
`List of people involved.`
`List of organisations involved.`
`List of means.`
One may see that this group answers the seven questions of situation awareness. The second group consists of the labels required to train supervised AI algorithms.

`Action type.`
`Parent event.` This label applies to social media re-posts. Set empty for original reporting.
`Mask.` Points out which fields are missed or distorted purposefully.
`Actual attack or noise event.`
The labelling system is designed to allow generated detests being used for actual attack detection and identification of the attack or operation type. Also, four labels validate if defenders AI can distinguish actual events from re-posting and reconstruct actual events from numerous incomplete messages.

5. DISCUSSION

By designing the generator, the authors bridge the gap between human understanding and perception of planning military and hybrid operations and using AI-powered technologies. During the formulation of the model, this was referred to as bridging the gap between military art influenced by the commanders' training and experience, and the implementation of formal methods.

One of the significant limitations of the proposed generator is the reliance on human experts to provide the rules required for planning each of the steps. To provide such rules, a suitable experienced operator is expected to possess both sufficient knowledge of mission planning (military of hybrid) and the principles of how AI can generate appropriate data sets. While human involvement in the loop is a necessity, this may cause specific problems that must be overcome Jenkins et al. (2012), Stanton et al. (2009).

The dataset's structure assumes that it would be used to train and validate AI systems with minimum possible data pre-processing. This makes it impossible to use generated data sets to validate more complex systems, including data scarping and text data analysis.

6. CONCLUSIONS

Limitations of human beings to perceive and process information lead to the need to use AI-assisted systems for the early recognition and modeling of hybrid and military operations. This contribution describes the process of synthetic data set generation to train and validate the use of AI-based systems in the early detection of military and hybrid operations. There are directions for the future research. The first is to investigate cooperation between human actors and AI-based systems supporting their activities to identify possible conflicts, in the same way as proposed by Vanderhaegen (2021). The second one may seem more obvious and will be directed towards stricter formalizing the concepts and embedding the generated into the closed-loop simulation systems.

ACKNOWLEDGEMENTS

This work was partially supported by the European Union European Regional Development Fund through the grand for Domain independent semi-automatic data model and relation logic for decision support. This work in the project ”ICT programme” was partially supported by the European Union through the European Social Fund.

REFERENCES

Bahşi, H., Nõmm, S., and La Torre, F.B. (2018). Dimensionality reduction for machine learning based iot botnet detection. In *2018 15th International Conference on Control, Automation, Robotics and Vision (ICARCV)*, 1857–1862. IEEE.

Balaban, M. and Mielniczek, P. (2018). Hybrid conflict modeling. In *2018 Winter Simulation Conference (WSC)*, 3709–3720. IEEE.

Cayirci, E., Bruzzone, A., Longo, F., and Gunneriusson, H. (2016). A model to describe hybrid conflict environments. In *13th International Multidisciplanary Modeling & Simulation Multiconference (I3M 2016), 26-28 September 2016, Larnaca, Cyprus*, 52–60. CAL-TEK Srl.

Cooper, C. and Zito, M. (2007). Realistic synthetic data for testing association rule mining algorithms for market basket databases. In J.N. Kok, J. Koronacki, R. Lopez de Mantaras, S. Matwin, D. Mladenič, and A. Skowron (eds.), *Knowledge Discovery in Databases:*

PKDD 2007, 398–405. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hudson, J.B. (1998). Intelligence officer handbook. Technical report, Federation of American Scientists. URL `https://irp.fas.org/doddir/army/fm34-8-2.pdf`.

Huszár, F., Ktena, S.I., O'Brien, C., Belli, L., Schlaikjer, A., and Hardt, M. (2022). Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119(1), e2025334119. doi:10.1073/pnas.2025334119. URL `https://www.pnas.org/doi/abs/10.1073/pnas.2025334119`.

Infantry Training Screw (2019). The non-combat, combat estimate. URL `https://wavellroom.com/2019/09/03/the-non-combat-combat-estimate/`.

Jenkins, D., Walker, G., Rafferty, L., Revell, K., Stanton, P., Salmon, P., Harris, P., and Salas, E. (2012). *Digitising Command and Control: A Human Factors and Ergonomics Analysis of Mission Planning and Battlespace Management*. Human Factors in Defence. Ashgate Publishing Limited. URL `https://books.google.ee/books?id=s996BgAAQBAJ`.

Kott, A. and Corpac, P.S. (2007). Compoex technology to assist leaders in planning and executing campaigns in complex operational environments. Technical report, DEFENSE ADVANCED RESEARCH PROJECTS AGENCY ARLINGTON VA.

Lieberman, S. (2012). Extensible software for whole of society modeling: Framework and preliminary results. *Simulation*, 88(5), 557–564.

Meidan, Y., Bohadana, M., Mathov, Y., Mirsky, Y., Breitenbacher, D., Shabtai, A., and Elovici, Y. (2018). N-baiot: Network-based detection of iot botnet attacks using deep autoencoders. *IEEE Pervasive Computing*, 13(9).

Omari, A., Langer, R., and Conrad, S. (2008). Tartool: A temporal dataset generator for market basket analysis. In C. Tang, C.X. Ling, X. Zhou, N.J. Cercone, and X. Li (eds.), *Advanced Data Mining and Applications*, 400–410. Springer Berlin Heidelberg, Berlin, Heidelberg.

Skaza, J. (2015). Mathematical modeling of trending topics on twitter. In *Bryant University, Honors thesis*.

Sokolowski, J.A. and Banks, C.M. (2007). From empirical data to mathematical model: using population dynamics to characterize insurgencies. In *Proceedings of the 2007 Summer Computer Simulation Conference*, 1120–1127.

Sokolowski, J.A. and Banks, C.M. (2009). *Modeling and simulation for analyzing global events*. John Wiley & Sons.

Stanton, N., Walker, G., Jenkins, D., and Salmon, P. (2009). Ergonomics challenges for digitisation of mission planning systems. 300 – 309. URL `https://www.scopus.com/inward/record.uri?eid=2-s2.0-84859905215&partnerID=40&md5=c85bda2cb081b4cc0fa59836f711fd20`.

Thiele, R.D. (2015). The new colour of war–hybrid warfare and partnerships. *World Politics of Security. Rio de Janeiro: Konrad Adenauer Foundation*, 47–59.

Unknown group of authors (2018). Military writer handbook. Technical report, Royal Military College of Canada. URL `https://moodle.rmc.ca/dcs/DCE080/Content/ModulePages/Module_01_5W.htm`.

Vanderhaegen, F. (2021). Heuristic-based method for conflict discovery of shared control between humans and autonomous systems - a driving automation case study. *Robotics and Autonomous Systems*, 146, 103867. doi:https://doi.org/10.1016/j.robot.2021.103867. URL `https://www.sciencedirect.com/science/article/pii/S0921889021001524`.