Digital Streets of Rage: Identifying Rhizomatic Extremist Messages During a Hybrid Media Event using Natural Language Processing

Teija Sederholm, Petri Jääskeläinen, Milla Lonka and Aki-Mauri Huhtinen
Department of Leadership and Military Pedagogy, National Defence University, Helsinki, Finland

teija.sederholm@mil.fi
jaaskelainenpet@gmail.com
milla.lonka@outlook.com
aki.huhtinen@mil.fi

Abstract: This research explores how to identify extreme messages during a hybrid media event happening in a small language area by utilizing natural language processing (NLP), a type of artificial intelligence (AI). A hybrid media event gathers attention all sides of the media environment: mainstream media, social media, instant messaging apps and fringe communities. Hybrid media events call attention for participation and activities both in the physical world and online. On the darker side of media events, the media landscape can act as a channel for all kinds of disinformation, hate speech and conspiracy theories. In addition, fringe communities such as 4chan also spread hate speech and duplicated content during hybrid media events. From theoretical point of view, this connection between the physical world and information networks can be seen as rhizomatic in nature, because information spreads without regard to a traditional hierarchy. The result is that when individuals participate in a big media event, there is a viral awareness of different viewpoints and all kind of topics may be posted online for discussion. In addition, in rhizomatic context different kind of arguments can twist each other, "copy and paste", and create very diversity meanings of new comments. The role of extremist speech in online spaces can have effects in physical world. The focus of this paper is to present the findings of a case study on messages posted online by three different actor groups who participated in demonstrations organized on Finnish Independence Day. In this research, two data sets were collected from Twitter and Telegram and Natural Language Processing (NLP) was used to classify messages using extremist media index labels. Three actor groups were identified as participating in the demonstrations, and they were labelled as: far-right, antifascists and conspiracists. Computational analysis was done by using NLP to categorize the messages based upon the definitions provided by the extremist media index. The analysis shows how AI technology can help identifying messages which include extremist content and approve the use of violence in a small language area. The model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible. This article is part of larger project related to extremist networks and criminality in online darknet environments.

Keywords: Extremism, Social Media, Artificial Intelligence, Rhizome

1. Introduction

The media landscape acts as a channel for all kinds of disinformation, hate speech and conspiracy theories. During big media events it is not uncommon for demonstrations to be coordinated by different groups using online resources. Over the last ten years, anarchists, far-right and antifascists have organized demonstrations at events surrounding The Independence Day of Finland. However, at the same time, demonstrations happen also in hybrid media environment including e.g., social media platforms, instant messaging apps and the mainstream media (Sumiala et al, 2018). From a theoretical point of view, this connection between streets and information networks can be seen as a rhizomatic nature: "any point of a rhizome can be connected to any other" or "any point of a rhizome can be fragmented to any other" (Deleuze and Guattari, 1980). The rhizomes show how topics appear and twist the discussion related to the hybrid media event.

The purpose of the research is to study the issues and possibilities of identifying Finnish extremist messaging with natural language processing (NLP) during a large Finnish hybrid media event, the Independence Day. While most of the discussion during a hybrid media event is not extremist, the spreaders of extremist media content can attach to the discussion in online space. The concept of rhizomes (Deleuze and Guattari 1980; 1988) is used to understand this phenomenon. The artificial intelligence is taught with messages labelled with the three-level Extremist media index (Holbrook, 2015). The aim of this research is to show possibilities of using combination of NLP and human analysis in identification of content endorsing violence.

A hybrid media event focuses on the contribution that different actors, messages and platforms make to a media event. These components create hybrid interplay in a complex network between mass media outlets, internet-based platforms and mobile communication technologies. A hybrid media event flows fluently between different actors and platforms creating an event that exists in multiple dimensions at once, connecting through a variety of ways. A hybrid media event consists of five elements; actors, affordances, attention, acceleration and affect. Actors can be both human actors and the platforms themselves. Affordance refers to the way different platforms

can be used for different purposes by different operators. Attention is crucial for getting more exposure for the content. Affects are means to accumulate attention and manage meanings such as strong emotions like fear or hate. Acceleration and movement are critical for hybrid media events, since they cause the content to circulate wider. These five elements interact with each other as a rhizomatic model, and create a basis for understanding how the hybrid media environment works. (Sumiala et al, 2018).

Actors with visibility on social media, e.g. accounts of mainstream media, direct much of the attention during a hybrid media event (Uusitalo and Valaskivi, 2020), though other actors in social media play a significant role as well. In addition, fringe communities, such as the imageboard website *4chan,* can be used to spread hate speech and duplicated content to gather attention (Zelenkauskaite et al, 2020). In addition, fringe communities have a significant role in circulating disinformation and conspiracy theories (Zannettou et al, 2017). Conspiracy theories, disinformation and hate speech can be used by different actors as a part of extremist messaging to dehumanize their 'enemies' and gather attention in hybrid media environment (Stecula and Pickup, 2021).

Hybrid media events can be described with a concept of a *rhizome.* Deleuze and Guattari (1980; 1988) use the term *rhizome* to describe a model of multiple, non-hierarchical entries and exit points in data representation and interpretation. A rhizome is an opposite to an arborescent (hierarchic, tree-like) conception of knowledge working with dualist categories and binary choices. Instead, a rhizome works with flat connections, and at any point of a rhizome it can be connected to any other part of the rhizome. Therefore, a rhizome may be broken at any point, and start up again later in some other lines. Thus, it is not possible to create any structural models, because a rhizome resembles more of a map without tracking. Due to this, actors may not know the end state of their actions and thus messages can have unpredictable but vital effects. Hence, it can be difficult to make differences between threats and opportunities in advance in a rhizomatic environment (Deleuze and Guattari 1988.)

The rhizomatic structure between the extreme and 'normal' content creating and sharing during a hybrid media event could be explained with a metaphor of Lego bricks versus watercolour: In principle, it is possible to extract a brick from a Lego-house without influencing its structure (Röck 2019) whereas in watercolour painting there is no exact or simple way to separate one colour from another afterwards. It is possible to extract the distortions, in this research the extremist media content, from the information flow. However, they stay connected to the normal discussion via rhizomes like colours in a watercolour painting. The rhizomatic nature of a hybrid media event helps those distortions to receive more attention. The actors participate the hybrid media event by commenting, sharing and reacting to events in both physical and digital world. Different actors can use conspiracy theories or mis- and disinformation to spread their own agenda, while attending a discussion about an ongoing topic (Sederholm, Huhtinen and Jääskeläinen, 2020). In this study, the groups participating the demonstrations are resistance movements trying to influence narratives created by officials and mainstream media. The four principles of rhizome Connection, Fragmentation, Escaping presumptions and Main role are used to categorize the extremist media content data.

Extremism could be characterized as an "ideological position embraced by those anti-establishment movements, which understand politics as struggle for supremacy rather than as peaceful competition between parties with different interests seeking popular support for advancing the common good" (Bötticher, 2017). In an extremist worldview, there are only friends and foes. The role of extremist speech in online spaces can have effects in the physical world. For example, online hate speech that include far-right political views can predict offline hate crimes against minorities (Gallacher and Heendric, 2021). This study examines three extremist groups; the far-right, antifascists and conspiracists. The distinctions between terms of right-wing extremism and right-wing radicalism are not clear even between academics, but both include authoritarianism, their anti-democracy, and their exclusionary and/or holistic nationalism with accompanying characteristics of xenophobia, racism, and populism (Carter, 2018). *Conspiracy theories are defined with a conspiracy narrative, that consists of a "group of malevolent people, conspiring to bring about some state of affairs to the detriment of the people in general", and this narrative does not exist outside of the conspiracy theory (Konda, 2019).* The view of antifascists as extremists has been prevailing in recent years around the world. Most of antifascism is seen nonviolent. In the US, the antifascists exercise restraint from violence, but "even if most of their 'everyday anti-fascism' is non-violent, militant anti-fascists use violence in confronting targets" (Copsey and Merill, 2020).

2. Data & Methodology

Two datasets were collected for this research. One from Twitter, consisting of three lists of words identified by researchers beforehand. The lists of words were determined based on the words likelihood to be used in content

related to actors participating in demonstrations. Second dataset was gathered from Telegram, and it consists of ten groups identified as potential places for extremist messaging. These datasets represent two aspects of the hybrid media event: a public social media platform and a closed instant messaging app. The datasets were collected from one day before to one day after the Finnish Independence Day demonstrations. The data was labelled by teaching NLP artificial intelligence to label Finnish messages with extremist media index.

The focus of this paper are the three actor groups participating in demonstrations during the Finnish Independence Day: far-right, antifascists and conspiracists. The level of extremist messages was identified by using the extremist media index (Holbrook, 2015). The index consists of three levels: 1. Moderate, 2. Fringe and 3. Extreme. Level 3 has three sublevels which were excluded from this research. Extreme includes justifications of violence. In fringe, the content must be radical and have hostility towards outgroup, but without justifications of violence. This includes content which directly targets the group by aggressive intention or using names that go beyond traditional political discourse (e.g. racial slurs).

In Finland, about 25% of ages 15-74 uses Twitter weekly (Pönkä, 2021). It is not nearly the most popular platform, but in this paper, it represents a platform with active, easily followed discussions about a hybrid media event taking place. The Twitter data was framed with key words used in Finnish Twitter discussions about the participating groups. The words were categorized as demonstrations, far-right, antifascist, conspiracy theorist and other (Table 1). On Twitter, the data was gathered via Twitter API by searching tweets with a word list including words for the demonstrations and words the actors were likely to use for spreading extremist media content about their opposing sides.

Table 1: Twitter search term examples.

Word group	Description	Search term examples on Twitter
Demonstrations	Referring to different demonstrations and overall words related to demonstrations	demonstration, demonstrator, protest, protestors, march, Helsinki without Nazis, 612, Finland wakes up
Far-right	Referring to far-right actors	far-right, extreme right, nazi, national socialism, national socialist, nationalist, nationalistic
Antifascist/left	Referring to antifascist actors	extreme left, anarchist, woke, leftie, communism, communist, commie, socialism, socialist, antifa, antifascist, antifascism
Conspiracy theorists	Referring to people that believe in conspiracy theories	conspiracy theorist, conspiracy, tin foil hat, denialist
Other	Words that may escalate extremist speech during demonstrations	terrorism, terrorist, terror, terrorist attack, elite, criminal, immigrant, Soviet Union, violence

Among these search terms are some words, which may have a negative connotation that can influence the data. For example, “troll” may refer to real trolls or be a way of downplaying a message by saying it was sent by a troll. The role of the word list “Other” is to bring out potential extreme speech which occurs when mentioning civil movements, terrorism or certain groups of people as a rhetoric tool for spreading extremist ideology. The retweets without quotations were not included to this data, since it would multiply the amount of the actual content. The amount of the final data was 5600 tweets.

On Telegram, the data was gathered from nine groups created for discussion related to far-right, antifascists and conspiracy theorists plus one for overall political discussion. The search was limited to known groups consisting of the users likely to attend the demonstrations and discuss about them, plus a single group for overall political discussion (Table 2). The number of messages in all groups together was 1626.

Table 2: Telegram groups.

Telegram group type	Description of group	Telegram group name (names translated from Finnish)
Far-right	Relation to known far-right movements	Defend Finland, *Backstage*, Soldiers of Odin
Antifascist/left	Relation to known antifascist movements	Helsinki without Nazis, A-group, Antifascist metropolitan area
Conspiracy theorists	Relation to known conspiracy theory activists	Truth seekers discussing, The Future of Finland, Universal Superhero Academy
Other	Group for overall political discussion	Politics

The far-right data was *collected from three different Telegram groups. Defend Finland is a large public discussion channel for activists from multiple groups whereas Backstage is a smaller offshoot group of Defend Finland with*

fewer actors. Finnish Soldiers of Odin's is a chat group. The chosen conspiracist groups were "Truth seekers discussing"*, "Universal Superhero Academy" and "*The Future of Finland"*. They were the most active channels* and were thus identified as places for conspiracy theory discussion.

Finnish antifascists groups found on Telegram were channels used to spread information and not as much to discuss about events. This makes them very different to the other two other categories. We collected the messages from three channels, "Antifascist metropolitan area", "A-group" and "Helsinki without Nazis", since they were likely to spread content related to the demonstrations. There are most likely more Finnish antifascist channels or groups on Telegram, but these are the only public ones we could find.

There were some preconditions used in labelling the data. Content that says something negative without real hostility towards the outgroup, e.g. "the stupid leftists are ruining this country" is excluded from the category fringe. Other messages excluded from fringe are ones where outgroups are used as slurs either towards the individuals of ingroup, since they do not target the actual groups of people. Demonstrations are confrontational by default, which is why these preconditions are used to identify fringe and extremist speech in the context of demonstrations. The distinction between concrete threat of violence and users using violent speech as a metaphor can also be difficult to understand, both for human actors as well as the artificial intelligence. Actors saying "throw Nazis to the ocean" are using a common phrase without a concrete threat of violence. They likely don't mean the messages as a concrete action, but as a locution inside their group. In this paper the solution was to label known catchphrases, such as this, as fringe. In cases where the catchphrase is not commonly used in Finnish activism, this sort of speech was identified as extreme, since it can be qualified as a vague reference to violence. Implicit and vague references to violence were also labelled as extreme.

The computational analysis was done by teaching AI technology created and provided by Aiwo (www.aiwo.ai). One researcher used the extremist media index to label messages collected from both Twitter and Telegram. Text analysis applies novel Natural Language Processing (NLP) technology called Bidirectional Encoder Representations for Transformers (BERT), which is based on deep learning neural networks (Devlin et al, 2019). The methodology introduces common language models that are pre-trained, typically using large corpus of data in context of a selected language. The models are then fine-tuned for specific analysis tasks.

The BERT method represents text as a set of tokens that describes words or pieces of words, their context and order in sentences. Furthermore, the method encodes tokens to vectors which are then fed to a neural network. The training algorithm presents the same data vectors as input and output to the network so that some of the tokens are masked in input, which adapts the neural network to learn the masked tokens in their sentence context. This training process produces a common language model. In analysis tasks the classification layer is attached to the pre-trained language model, and this full neural network is fine-tuned using training data that consists of sample texts and their class labels. (Devlin et al, 2019)

This study used the English pre-trained BERT model (Hugging Face, 2018). For classification, 1698 samples of the tweets were labelled from Twitter to three classes: extreme (75 samples), fringe (347 samples) and moderate (1276 samples). This data was sampled randomly to training and validation sets which were used to fine-tune the classification model and to evaluate the resulted model. The model performance was evaluated using F-score to obtain classification accuracy of the validation data.

Table 3. F-score of validation set with three classes.

Label	True positive	True negative	False positive	False negative	F1
Moderate	129	56	18	9	0.905263
Fringe	42	143	11	16	0.756757
Extreme	5	191	5	11	0.384615

Table 4. F-score of validation set with two classes.

Label	True positive	True negative	False positive	False negative	F1
Moderate	132	27	3	7	0.963504
Fringe/Extreme	27	132	7	3	0.843750

The first approach classified tweets to three classes: moderate, fringe and extreme. Quantitatively, F-score presented in Table 3 shows that the classification has rather large portion of false hits, especially with the extreme class. In the next approach the fringe class was merged to the extreme class which led significantly better result (see Table 4). This implies to two causes. First, the number of extreme class is too small compared other classes, and secondly, the labelling of the fringe class is potentially too ambiguous with the extreme class.

3. Findings

Based on the scores, the model recognizes moderate messages well and is able to differentiate it from other categories, but had issues with separating fringe and extreme content from each other. This issue rose from the amount of data, which was not as large as it should possibly be for teaching the NLP-technology. Even when the data was collected during a hybrid media event with extremists participating in it, they weren't active enough in social media to give sufficient data for modelling.

Based on the computational analysis made by the AI technology, the number of extreme messages was very low and the number of fringe messages was almost as big as the number of moderate messages. The reason for this is that the collection of data was based on controversial and emotion-stirring topic of demonstrations during the Independence Day. What is noticeable is that even if the number of messages was largest during the Independence Day, the relative amount of fringe messages was larger after the event. This may be explained by users reacting to the media event not only in real-time, but after it. The hybrid media environment expands the possibilities of going back in time and space inside the digital environment, which is why the users will react to older topics. The same phenomenon occurred with extreme messages.

The far-right groups had the most active discussion but much of it was due of one provocative actor acting before and after the demonstrations. The use of conspiracist and antifascist chats and groups were different, with both acting mainly as a place to share information in. The actors inside conspiracist chat groups perceived authorities as their outgroup but didn't use extremist messaging to talk about them. The antifascist groups didn't mention their outgroups at all.

The AI was taught to categorize messages as fringe and as extreme from each topic. Through the rhizomatic model, the AI shows what kind of role these messages had during the hybrid media event. The AI identified 11 main topics from the data: Parties and politics, Security, Independence Day celebrations, Society, Citizens, Governance, Traditions and culture, Health security, Relations to outside world, history of Finland and the experience of being a Finn. All of these topics were broken to smaller subtopics. To compensate the fact that the AI technology was not able to distinguish extreme and fringe properly, another two analyses were done with the focus on them. This way the computational method was able to create overall picture of the extreme and fringe messages. Tables 5 and 6 show the top three themes (neo-nazism, extremism, fascism) based on the amount of fringe and extreme messages. They were analysed with their rhizomatic connections.

Table 5. Fringe messages analysed with rhizomatic connections.

Rhizome	Topic: Neo-nazism	Topic: Extremism	Topic: Fascism
Connected	Against Nazism and use of term "Nazi" to describe outgroup regardless of their nazism	Against all extremist groups	Discussions against far-right demonstrators and framing of antifascists as fascists
Fragmented	Users using Nazism to frame their (political)outgroup	Use of "whataboutism" to move discussion away from far-right topics	Break the discussions about far-right with framing antifascists as real fascists
Escape presumptions	Messages calling pro-left and governmental actors Nazis	No mentions of conspiracy theorists	Discussions about antifascism as real fascism
Main role	To oppose the far-right demonstrations	To oppose all demonstrations, though the focus is on framing antifascists as bad as far-right	To oppose far-right and all extreme actors

Table 6. Extreme messages analysed with rhizomatic connections.

Rhizome	Topic: Neo-nazism	Topic: Criminality	Topic: Terrorism
Connected	Against the far-right	Accusations of governmental actors being criminals	Discussions about a plotted far-right terrorist plot and its connections

Rhizome	Topic: Neo-nazism	Topic: Criminality	Topic: Terrorism
Fragmented	Shift the focus of discussions to threat of far-right or the antifascists being 'real Nazis'	Discussions mainly unrelated or vaguely related to demonstrations	To shift the focus into the role of political right in supporting terrorists
Escape presumptions	Some messages using Nazism to describe political left and governmental actions	One individual spamming multiple messages written in same format. This actor also wrote much fringe content.	Some actors accusing antifascists as terrorists based on the situation in USA
Main role	To diminish the human value of far-right and their right to action in society	To create and spread distrust in authorities	To spread accusations of potential extreme actors and their role in supporting terrorism

Based on the analysis, it's possible to teach NLP to categorize Finnish-language messages during a hybrid media event and identify which of those include extreme ideas. The extreme does not automatically mean the messages incite users for violence, but they can be used to justify the use of violence during demonstrations.

4. Discussion

The purpose of this research was to study the issues and possibilities of identifying Finnish extremist messaging with natural language possessing (NLP) during a large hybrid media event, the Finnish Independence Day. The data was gathered from two social media platforms, Twitter and Telegram. Overall, the demonstrations during the Independence Day were mostly peaceful and the number of participants was lower than expected. However, it can still be argued through analysed data that the model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible during a hybrid media event.

The results indicate that it is possible to teach NLP to categorize Finnish-language messages during a hybrid media event and to identify extremist text content. It is notable, that extreme content does not automatically mean the messages incite users for violence, but they can be used to justify the use of violence during demonstrations. The identification of the differences between moderate and fringe/extreme messages was very accurate, but the finetuning between fringe and extreme messages was not as successful. As Holbrook (2015) writes, "There are grey areas between categories and some boundaries are difficult to define, particularly between vague fringe material and vague level 1 extreme material therefore a degree of overlap between categories should be assumed". This is most probably due to Finnish language and a low number of extreme messages. It is possible, that a relatively larger data could make the recognition model more accurate.

There can be multiple reasons for the low number of extreme messages, even when part of the data was collected from Telegram groups associated with potential extremist movements. Firstly, the chat groups were publicly known and this may have influenced the nature of discussion. Secondly, the participants were on the streets during the event or celebrating Independence Day elsewhere and thus not so active on social media. Lastly, there was no outgroup present to target with extreme messages. On Twitter, the conversation was more heated with the outgroups being present on the same platform. Also, the role of antifascist Telegram-channels was to send information, not embrace a discussion. There are still some issues to be solved. How to recognize the difference between hostility and pungent critique towards an outgroup? The outgroups such as Nazis and communists are used both to discuss the groups and to stigmatize the actors. It is critical to teach the AI to recognize fringe and extreme messages. Another issue are words with negative connotations, like being able to distinguish the confrontative attitude towards communists from messages of what communists have done historically.

The model of rhizome was valid in making the connections between fringe, extremist content and moderate discussion visible during a hybrid media event. The use of the rhizome as model of data analyse reflects its expansive and ephemeral nature, rather than one which forecloses its conceptual possibilities, is promoted (Lawley, 2005). The extreme flows and blends into non-extreme discussion via rhizomes and gets mixed up in normal political discussions. The concept of rhizome can help to see the connections between extreme and moderate discussions, possibly to even predict where extreme and fringe messages may appear.

The messages analysed were collected afterwards meaning Twitter or Telegram did not label them breaking the terms of the platform. This might mean their ability to identify extremist messages in small language areas such as Finnish is inefficient. In addition, this means that these platforms still allow extremist content to spread even though it breaks their terms of use. If the social media cannot distract the colours of extreme from the watercolour painting of political discussion, it continues to influence radical actors and stir up emotions

regarding controversial topics, such as demonstrations and activism. For future research, the Holbrook's extremist media index can be used as a basis for the learning process for the natural language processing of extremist messaging in small language area.

Acknowledgements

This study is part of a research project funded by the Academy of Finland

References

Bötticher, A. (2017) Towards Academic Consensus Definitions of Radicalism and Extremism. *Perspectives on Terrorism*, *11*(4), 73–77. Available at: http://www.jstor.org/stable/26297896. (Accessed January 19, 2023).

Carter, E. (2018) Right-wing extremism/radicalism: reconstructing the concept, *Journal of Political Ideologies*, *23*(2), 157-182.

Deleuze, G. and Guattari, F. (1988) *A Thousand Plateaus Capitalism and Schizophrenia*, Bloomsbury Academic.

Deleuze, G. and Guattari, F. (1980) A Thousand Plateaus, Trans. Brian Massumi. London and New York: Continuum, 2004. Vol. 2 of Capitalism and Schizophrenia. 2 vols. 1972-1980. Trans. of *Mille Plateaux*. Paris: Les Editions de Minuit.

Devlin, J., Chang M-W., Lee K. and Toutanova K. (2018) BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*, NAACL-HLT 2019: Minneapolis, MN, USA*, Volume 1: 4171-4186

Gallacher, J. and Heerdink, M. (2021) Mutual radicalisation of opposing extremist groups via the Internet, *Preprint* doi:10.31234/osf.io/dtfc5

Holbrook, D. (2015) "Designing and Applying an 'Extremism Media Index'", *Perspectives on Terrorism*, 9(5), 57–68. Available at: http://www.jstor.org/stable/26297434 (Accessed January 19, 2023).

Hugging Face (2018) *BERT base model (uncased)*. Available at: https://huggingface.co/bert-base-uncased (Accessed May 15, 2022).

Jalonen, H., Paavola, J., Helo, T., Sartonen, M. and Huhtinen, A.-M. (2016) Understanding the Trolling Phenomenon: The Automated Detection of Bots and Cyborgs in the Social Media, *Journal of Information Warfare,* 15. 100-111.

Konda, T. M. (2019) *Conspiracies of Conspiracies: How delusions have overrun America,* The University of Chicago Press, Chicago, ISBN-13: 978-0-226-58576-5. doi:10.7208/chicago/9780226585932.001.0001

Lawley, S. (2005) Deleuze's Rhizome and the Study of Organization: Conceptual Movement and an Open Future*, Journal of Critical Postmodern Organization Science,* Vol 3 (4), 36-49. Available at: https://www.academia.edu/download/35484409/Tamara.pdf (Accessed January 19, 2023).

Leppänen, M. (2019) Poliisi: 612-kulkueessa oli 1 500 osallistujaa, äärioikeiston vastustajien mielenosoituksessa 2 700, kiinni otettiin 13 ihmistä, *Yleisradio* 6.12.2019. https://yle.fi/uutiset/3-11106587 (Accessed at November 29, 2021).

Pönkä, H. (2021) Sosiaalisen median katsaus 04/2021. Available at: https://speakerdeck.com/hponka/2021 (Accessed January 19, 2023).

Roozenbeek, J. and van der Linden, S. (2021) Inoculation Theory and Misinformation. *Published by the NATO Strategic Communications Centre of Excellence*, ISBN: 978-9934-564-49-9. Available at https://stratcomcoe.org/pdfjs/?file=/publications/download/Inoculation-theory-and-Misinformation-FINAL-digital-ISBN-ebbe8.pdf?zoom=page-fit (Accessed January 19, 2023).

Sederholm T., Jääskeläinen P. and Huhtinen A-M. (2021) Coronavirus as a Rhizome – The Pandemic of Disinformation, *International Journal of Cyber Warfare and Terrorism.* 11(2): 43–55. doi: 10.4018/IJCWT.2021040104

Stecula D. A. and Pickup M. (2021) Social Media, Cognitive Reflection, and Conspiracy Beliefs, *Front. Polit. Sci.* 3:647957. doi:10.3389/fpos.2021.647957

Starbird, K., Maddock, J., Orand, M.; Achterman, P. and Mason, R. M. (2014) Rumors, False Flags, and Digital Vigilantes: Misinformation on Twitter after the 2013 Boston Marathon Bombing*, In iConference 2014 Proceedings* (p. 654 - 662). doi:10.9776/14308

Uusitalo, N., and Valaskivi, K. (2020) The Attention Apparatus: Conditions and Affordances of News Reporting in Hybrid Media Events of Terrorist Violence, *Journalism Practice*. doi: 10.1080/17512786.2020.1854052

Zannettou, S., Caulfield, T., & De Cristofaro, E., Kourtellis, N., Leontiadis, I., Sirivianos, M., Stringhini, G. and Blackburn, J. (2017) The Web Centipede: Understanding How Web Communities Influence Each Other Through the Lens of Mainstream and Alternative News Sources, *Proceedings of the 2017 internet measurement conference*, 405-417. doi:10.1145/3131365.3131390

Zelenkauskaite, A., Toivanen, P., Huhtamäki, J. and Valaskivi, K. (2020) Shades of hatred online: 4chan duplicate circulation surge during hybrid media events, *First Monday*, *26*(1). doi: 10.5210/fm.v26i1.11075

An Analysis of Critical Cybersecurity Controls for Industrial Control Systems

Nkata Sekonya, Siphesihle Sithungu
University of Johannesburg, Johannesburg, South Africa

nkatasekonya61@gmail.com
siphesihles@uj.ac.za

Abstract: Industrial Control Systems (ICS) comprise software, hardware, network systems, and people that manage and operate industrial processes. Supervisory Control and Data Acquisition Systems (SCADA) and Distributed Control Systems (DCS) are two of the most prevalent ICS. An ICS facilitates the effective and efficient management and operation of industrial sectors, including critical infrastructure sectors like utilities, manufacturing, and water treatment facilities. An ICS collects and integrates data from various field controllers deployed in industrial contexts, enabling operators to make data-driven decisions in managing industrial operations. Historically, ICS were isolated from the internet, functioning as part of air-gapped networks. However, the efficiency improvements brought about by the emergence of Information Technology necessitated a shift towards a more connected industrial environment. The convergence of Information and Operational Technology (IT/OT) has made ICS vulnerable to cyberattacks. Due to the crucial nature of the infrastructure that ICS manage, cyberattacks against ICS may cause critical infrastructure sectors to experience downtime. This may have a crippling impact on a country's well-being and essential economic activities. Given the proliferation of cyber warfare, cyberattacks against ICS are increasingly significant at present, as was the case during the 2015 attack on Ukraine's power infrastructure, which was successful in causing a blackout that affected over 200 000 persons. The threat actors used malicious software known as "BlackEnergy3", which was created to interfere with the regular operation of the ICS in charge of controlling electrical substations. This was the first known instance of malicious software causing blackouts. In response to increasing cyberattacks against ICS, the SANS Institute, in a whitepaper titled "The Five ICS Cybersecurity Critical Controls", present five critical controls for an ICS cybersecurity strategy. This paper discusses ICS and the increased convergence of IT and OT. The paper also outlines significant cyberattacks directed at ICS. The paper then follows an exploratory research methodology done in response to the Five ICS Cybersecurity Critical Controls to determine the state of ICS literature that can help ICS operators secure their environments in accordance with the framework. Additionally, the ICS Cybersecurity Critical Controls are mapped to the NERC CIP standards, which provide guidance on the security of the Bulk Electric System (BES) and associated critical assets in North America.

Keywords: industrial control systems, critical infrastructure, cybersecurity

1. Introduction

Industrial Control Systems (ICS) play an important role in managing complex industrial processes, including those in critical infrastructure sectors like utilities, manufacturing, and water treatment facilities. ICS comprise networked software and hardware systems that facilitate the management of complex industrial processes by enabling process automation, monitoring, and safety in industrial environments (Asghar and Zeadally, 2019). In the past, ICS were separated from traditional IT networks (Thomas and Chothia, 2020). However, the increased adoption of technologies brought about by the emergence of Information Technology has seen a shift towards a more connected ICS environment. As a result of the convergence of Information and Operational Technology (IT/OT), new cybersecurity issues have emerged (Miller et al., 2021). Remote access to ICS, for example, enables remote management of industrial processes and remote maintenance of industrial equipment. This, however, leaves the ICS environment vulnerable to cyberattacks. The successful breach of ICS can result in the interruption of industrial processes, physical equipment damage, human casualties, and devastating physical repercussions (Alladi, Chamola and Zeadally, 2020). This is demonstrated – for example – by the many attacks on Ukraine's electricity infrastructure, which have resulted in outages (Hemsley and Fisher, 2018). In response to attacks on ICS environments, the SANS Institute published a whitepaper presenting cybersecurity controls most critical for ICS environments. The controls are 1) an implementation of ICS-specific incident response plan(s), 2) the use of a defensible architecture, 3) ICS network visibility and monitoring, 4) secure remote access, and 5) risk-based vulnerability management. The objective of the proposed controls is to guide industrial organisations in creating or enhancing their ICS security program, better protecting ICS environments from cyberattacks. The main objectives and contributions of this paper are as follows: To recommend security controls that conform to the SANS Five ICS Cybersecurity Critical Controls after having conducted an exploratory examination of the present state-of-the-art literature in ICS security. To map the SANS Five ICS Cybersecurity Critical Controls to the NERC CIP standards used in the protection of the Bulk Electric System (BES) and associated infrastructure.

To accomplish the stated objectives, the paper is structured as follows: Section 2 discusses the ICS architecture and the most common types of ICS. Section 3 discusses the convergence of IT and OT and the cybersecurity concerns brought by this convergence. Section 4 discusses major cyberattacks affecting ICS infrastructure. The contribution of this work is Section 5, which examines and aligns the SANS Five ICS Cybersecurity Critical Controls with existing ICS security literature. Section 6 provides a mapping of the ICS Cybersecurity Critical Controls to the NERC CIP standards. Section 7 concludes the paper.

2. Industrial Control Systems

Industrial Control Systems (ICS) comprise networked software and hardware systems that allow for the control and supervision of complex industrial processes (Drias, Serhrouchni and Vogel, 2015). Two of the most common ICS are Distributed Control Systems (DCS) and Supervisory Control and Data Acquisition (SCADA) systems. A DCS enables distributed control within a defined geographic area, whereas SCADA systems are used for centralised control and monitoring of geographically scattered assets (Yadav and Paul, 2021). Building automation systems (BAS), physical access control systems (PACS), and safety systems are additional ICS (Stouffer et al., 2022). Common ICS and their components are covered in this section, including the sectors in which they are typically used.

2.1 SCADA Systems

SCADA systems centralise data acquisition and control of geographically scattered assets and are utilised in distribution systems such as water distribution, oil and gas pipelines, transportation systems, and electrical utility transmission (Stouffer et al., 2022). SCADA systems are made up of software, hardware, and network components that can be classified as 1) field devices, 2) communication systems, and 3) supervisory systems. Field devices comprise sensors and actuators for data collection and physical control, as well as Programmable Logic Controllers (PLCs) or Remote Terminal Units (RTUs) responsible for control logic and communication with the control server (Pliatsios et al., 2020). The use of radio, satellite, or power line-based communication mediums, among others, by communication systems allows for communication between field equipment and supervisory systems (Stouffer et al., 2022). Also included in the communication systems are ICS-specific communication protocols like Modbus and DNP3 (Asghar, Hu and Zeadally, 2019). Supervisory systems include the Master Terminal Unit (MTU), which transmits control data to field equipment and collects and stores data in the Data Historian (Drias, Serhrouchni and Vogel, 2015). Operator interaction with data from the MTU is made possible through the Human Machine Interface (Asghar, Hu and Zeadally, 2019). The design of a SCADA system is shown in Figure 1.

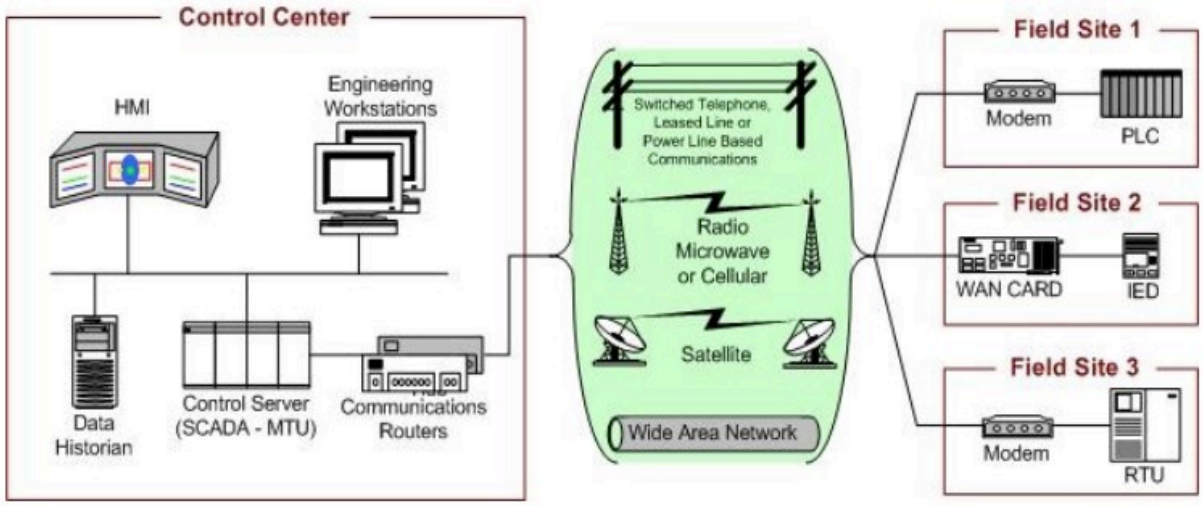

Figure 1: SCADA System General Layout (Stouffer, Falco and Scarfone, 2011)

2.2 Distributed Control Systems

Distributed Control Systems (DCS) are used in the management and control of industrial processes within the same geographic area (Asghar, Hu and Zeadally, 2019). A DCS has a supervisory level of control overseeing multiple controllers distributed across the industrial environment (Bolton, 2021). DCS are typically used in power generation plants, oil refineries and manufacturing (Stouffer et al., 2022). DCS employ most of the components discussed under SCADA systems, with PLCs acting as local controllers. Both SCADA and DCS fall under the

umbrella of Operational Technology. The next section discusses the convergence of Information and Operational Technology.

3. IT and OT Convergence

The Industrial Control Systems used to manage and operate industrial processes fall under Operational Technology (OT). Historically, ICS operated in an environment isolated from enterprise Information Technology (IT) networks and the public internet (Murray, Johnstone, and Valli, 2017). This provided a level of security because, to carry out a successful attack, an adversary had to have physical access to the ICS. While a deterrent, it should be noted that a motivated adversary, such as the threat actors behind Stuxnet, which targeted Iran's Natanz uranium enrichment facility, could still infiltrate such an environment isolated from the internet through infected USB sticks (Miller et al., 2021). Nonetheless, ICS, which traditionally operated in a relatively secure environment, are now connected to enterprise IT and the internet at large. This presents challenges as ICS are not equipped to deal with cybersecurity challenges associated with IT systems. It is in this context that this section discusses the convergence of IT and OT and the resultant cybersecurity challenges.

Operational Technology (OT) systems include SCADA, DCS and associated industrial communication networks. OT is primarily concerned with the safety and consistency of industrial processes, whereas Information Technology (IT) is concerned more with the management of data (Hahn, 2016). The ability to tolerate delays is the first of several distinctions between OT and IT environments that are described in (Stouffer et al., 2022). Because OT powers critical infrastructure sectors, it is less tolerant of delays than IT. In an industrial facility, a delay in the reaction of safety systems might have devastating implications. OT has stricter availability requirements than IT since downtime can have a cascading effect due to the interdependence of critical infrastructure driven by OT systems. Downtime in telecommunications infrastructure, for example, has a negative impact on businesses because the internet is at the heart of many corporate activities.

Another key distinction is that OT prioritises safety above security; as a result, security measures that may jeopardise the safety of OT systems are not put in place. OT operators place a higher priority on the reliability of the industrial process, whereas IT suggests that security controls and upgrades be implemented as soon as they are available. This creates difficulties since OT systems' vulnerabilities may not be fixed for a long time. Systems in OT contexts have fewer resources than their IT counterparts, making it challenging to implement security controls like encryption. The issue is made worse by the usage of proprietary protocols in OT systems, which differ from those used in IT. Applying security rules that are ubiquitous in IT systems becomes challenging or nearly impossible as a result. According to Stouffer et al. (2022), another problem impeding OT system security is reliance on a single OT provider. Third-party security services can invalidate vendor agreements, leaving enterprises operating OT systems with few alternatives for safeguarding their systems. This is especially problematic given that the lifespan of OT components might range from 10 to 15 years (Hahn, 2016).

Cyberattacks on OT environments have escalated, which is a major problem caused by the convergence of IT and OT systems. OT settings are now susceptible to attacks that are widespread in IT systems. This is particularly troubling because OT runs critical infrastructure, whose interruption might have a severe impact on a nation's economy, hinder its capacity to run or even put people's lives in peril. A subset of major cyberattacks targeted at OT systems is discussed in the following section.

4. Industrial Cybersecurity Events

Cyberattacks on Industrial Control Systems may interrupt the operation of critical infrastructure. This is demonstrated by the incidents mentioned in this section. Cyberattacks mentioned in this section are the most significant in terms of their ability to have physical repercussions, disrupt industrial activities, and cause human casualties. This section covers a selection of major ICS attacks.

4.1 Stuxnet

At the time, Stuxnet was the most sophisticated attack on critical infrastructure, and it was used to thwart Iran's nuclear program. In 2010, researchers discovered the Stuxnet malware, which was aimed at Iran's Natanz uranium enrichment complex. Because these facilities were not connected to the public internet, initial access is believed to have been gained through an infected USB stick (Miller et al., 2021). After obtaining access to the network, Stuxnet exploited vulnerabilities in conventional IT systems to infect the PLCs in charge of managing the facility's centrifuges. This allowed Stuxnet to alter the centrifuge speeds there, leading them to malfunction (Firoozjaei et al., 2022).

4.2 BlackEnergy3

In 2015, Ukraine's electricity infrastructure was attacked, resulting in power disruptions for about 200 000 persons (Alladi, Chamola and Zeadally, 2020). BackEnery3 malware was employed in the attack, which was suspected to have been sent using spear-phishing emails (Firoozjaei et al., 2022). The assault used the third version of the BlackEnergy malware, originally detected in 2007, and the KillDisk trojan to wipe out any evidence on targeted systems (Miller et al., 2021). This was the first recorded case of malicious software causing blackouts, and it served as a warning of how devastating an attack on the ICS running critical infrastructure can be.

4.3 Industroyer

In 2016, Industroyer malware, also known as the Crash Override malware, was utilized in another attack on Ukraine's power infrastructure. The initial access is said to have been obtained using spear phishing emails, which are proving effective (Hemsley and Fisher, 2018). This attack was successful in knocking off electricity to 20% of Ukraine's capital, Kyiv (Alladi, Chamola and Zeadally, 2020).

4.4 Triton

Triton is an attack framework that was used in an attack on an ICS at a Saudi Arabian petrochemical company. Triton was used to target a Safety Instrumented System (SIS) in charge of monitoring SIS safety controllers, which are the plant's final line of defence against disasters (Miller et al., 2021). SIS controllers are in charge of avoiding plant failures, and their disruption might have severe consequences for both personnel safety and the surrounding environment. The attack was successful in interrupting facility operations since industrial systems had to be shut down (Hemsley and Fisher, 2018).

Attacks on ICS can result in catastrophic outcomes, as the incidents covered in this section demonstrate. The next section discusses the SANS Five ICS Cybersecurity Critical Controls and looks at recent research to find security measures that can help ICS operators secure their environments in accordance with the framework.

5. SANS Five ICS Cybersecurity Critical Controls

In response to recent cyberattacks on Industrial Control Systems (ICS) around the world, ICS Cybersecurity experts Robert M. Lee and Time Conway from the SANS Institute introduced Five ICS Cybersecurity Critical Controls in their whitepaper titled “The Five ICS Cybersecurity Critical Controls”. The objective of the proposed controls is to guide industrial organisations in creating or enhancing their ICS security program (Lee and Conway, 2022). The proposed cybersecurity controls most critical to ICS systems are 1) an ICS incident response, 2) a defensible architecture, 3) ICS network visibility monitoring, 4) secure remote access, and 5) risk-based vulnerability management (Lee and Conway, 2022). This section discusses the proposed controls and explores the current ICS cybersecurity literature to identify, discuss and map methods that can guide ICS operators to secure their environments in accordance with the proposed framework.

5.1 ICS Incident Response

An incident response plan details the steps taken to respond to a cyber incident, which includes (but is not limited to) preparation, detection, containment, eradication, recovery, and post-incident activity (Staves et al., 2020). (Lee and Conway, 2022) advise that industrial organisations should have an ICS-specific incident response plan. Organisations are advised to begin by identifying scenarios that pose the most risk to their organisation. This process should be intelligence-driven, studying past successful attacks on organisations in the same sector. Secondly, organisations should identify high-consequence scenarios regardless of whether an attack has affected their sector in the past (e.g., ransomware attacks) and determine the steps an attacker would need to take to successfully compromise the organisation’s infrastructure. Lastly, organisations ought to perform tabletop exercises to test the maturity of their incident response plan.

The work of (Miller et al., 2021) provides a comprehensive list of publicly reported cyberattacks aimed at ICS. The study covers attacks which occurred as early as 1988 and as late as 2020. For each cyberattack, the authors discuss the method used by threat actors to gain initial access to the OT environment, the threat actors responsible for the attack, the sector of the targeted organisation and the impact of the attack. This study can assist ICS operators in identifying scenarios that pose the most risk to their organisation. Organisations can study cyberattacks that have in the past targeted their sector. Figure 2 presents a summary of cyberattacks aimed at ICS.

Attack	Date	Initial Access	Threat Actor	Sector	Impact
PLC Password Change	1988	Workstation Compromise	Insider	Manufacturing	Denial of Control
Ignalina Nuclear Power Plant	1992	Workstation Compromise	Insider	Civil Nuclear	Loss of Productivity and Revenue
Chevron Refinery Emergency Alarm System	1992	Workstation Compromise	Individual	Chemical	Loss of Productivity and Revenue
Salt River Project	1994	Internet Accessible Device	Individual	Energy and Water	Loss of Productivity and Revenue, Disk Wipe
Omega Engineering	1996	Workstation Compromise	Individual	Manufacturing	Disk Wipe
Worcester, MA Airport	1997	Internet Accessible Device	Individual	Transport	Loss of Productivity, Revenue, Availability, and Safety
Gazprom	1999	Unknown	Organised Group + Employee	Chemical and Energy	Loss of Productivity and Revenue
Bradwell Nuclear Power Plant	1999	Workstation Compromise	Insider	Civil Nuclear	Disk Wipe
Maroochy Water System	2000	Wireless Compromise	Insider	Water	Damage to Property
Cal-ISO System	2001	Unknown	Nation State	Energy	None Disclosed
Virus on Manufacturing System	2001	Spearphishing	Nation State	Manufacturing	Loss of Productivity and Revenue
Houston Port	2001	Internet Accessible Device	Individual	Transport	Loss of Productivity and Revenue
Gas Processing Plant	2001	Trusted Relationship	Supplier	Chemical	Loss of Productivity and Revenue
PDVSA	2002	Internet Accessible Device	Organised Group	Chemical	Loss of Productivity and Revenue, Disk Wipe
Flight Planning Computer	2003	Unknown	Individual	Transport	Loss of Productivity and Revenue
CSX Train Signalling System	2003	Spearphishing	Unknown	Transport	Loss of Productivity and Revenue
Contractor Infects SCADA Network	2004	Replication Through Removable Media	Unknown	Food	Loss of Productivity and Revenue
Daimler Chrysler Plants	2005	External Remote Service	Individual	Manufacturing	Loss of Productivity and Revenue
Tehama-Colusa Canal	2007	Workstation Compromise	Individual	Water	Damage to Property
Lodz Tram System Hacked	2008	External Remote Service	Individual	Transport	Loss of Safety
US Power Grid	2009	Internet Accessible Device	Nation State	Energy	None Disclosed
Hospital HVAC	2009	Workstation Compromise	Insider	Health	Loss of Safety
Night Dragon	2009	Exploit Public-Facing Application	Organised Group	Energy	Theft of Operational Data
Sality Virus Infects DVS Servers	2009	Unknown	Unknown	Chemical	Loss of View
Stuxnet	2010	Replication Through Removable Media	Nation State	Civil Nuclear	Damage to Property, Manipulation of View and Control
Shionogi	2011	Workstation Compromise	Individual	Health	Disk Wipe
Niagra AX	2012	Internet Accessible Device	Unknown	Manufacturing	Manipulation of Control
Espionage on Iranian CI	2012	Replication Through Removable Media	Nation State	Chemical	Theft of Operational Data, Unintentional Disk Wipe
Turbine Control System	2012	Replication Through Removable Media	Organised Group	Energy	Loss of Productivity and Revenue, Theft of Operational Data
Rye Brook Dam	2013	Internet Accessible Device	Organised Group	Water and Energy	None Disclosed
European Public Utility Services Attacked	2014	Spearphishing	Organised Group	Various	Denial of Service, Theft of Operational Data
German Steel Mill	2014	Spearphishing	Unknown	Manufacturing	Damage to Property
Ukrainian Energy	2015	Spearphishing	Organised Group	Energy	Loss of Productivity and Revenue
Ukrainian Energy	2016	Spearphishing	Organised Group	Energy	Disk Wipe, Loss of Productivity and Revenue, Loss of Safety
Wolf Creek	2017	Spearphishing	Organised Group	Civil Nuclear	None Disclosed
Cadbury Factory Attack	2017	External Remote Service	Organised Group	Food	Loss of Productivity and Revenue
Triton/Petro Rabigh	2017	Workstation Compromise	Nation State	Chemical	Denial of Control, Loss of Safety
Norsk Hydro	2019	Spearphishing	Unknown	Manufacturing and Energy	Loss of View
Triton/Undisclosed	2019	Workstation Compromise	Nation State	Undisclosed	Denial of Control, Damage to Property, Loss of Safety
Hackers Target Oil Producers	2020	Spearphishing	Unknown	Chemical	Theft of Operational Data
Israeli Water Facilities Attacked	2020	Internet Accessible Device	Organised Group	Water	None Disclosed
Cyber-Attack on Shahid Rajaie Port	2020	Unknown	Nation State	Transport	Loss of Productivity and Revenue
Honda Factories Cyber Attack	2020	Spearphishing	Unknown	Manufacturing	Denial of Control

Figure 2: Summary of ICS cyberattacks (Miller et al., 2021)

The work of (Firoozjaei et al., 2022) takes it a step further by analysing the most significant of these attacks and mapping the adversarial tactics and techniques against the MITRE ATT&CK framework. This is crucial, as it provides a mapping of the actions a threat actor would need to perform to successfully compromise an organisation's infrastructure. This enables the organisation to develop incident response plans specific to their ICS environments. Once scenarios posing the most risk to the organisation are identified, and appropriate incident response plans are in place, organisations should implement an architecture that allows them to detect and respond to the identified scenarios. The next subsection discusses the implementation of a defensible architecture.

5.2 Defensible Architecture

This control is concerned with establishing an architecture that satisfies the organisation's operational demands, is adaptive and robust, and supports the data gathering required for security operators to protect against the scenarios outlined in the first control (Lee and Conway, 2022).

Given the growing convergence of OT and IT, a common theme among network architectures is to provide segmentation between the IT and OT network. This is to ensure that attacks on IT do not propagate to the OT environment, affecting ICS. A unified architecture for ICS protection, proposed by (Zhou et al., 2020), offers integrated protection for networks, control systems, and physical processes. The design provides a secure network, secure ICS protocols, and secure end-to-end communications to guarantee the protection of sensitive process data. Securing control systems is done using a risk-based strategy that includes defences based on tolerance and prevention. Additionally, the design offers a defence against process-aware attacks. The next step after establishing a defendable architecture is to guarantee network visibility, which is covered in the next subsection.

5.3 ICS Network Visibility and Monitoring

One of the primary distinctions between Information and Operational Technology (IT/OT) is that IT prioritises data while OT prioritises processes. It is within this context that the visibility of the process in OT environments becomes crucial. Visibility in OT extends beyond individual systems to include the monitoring of ICS-specific protocols (Lee and Conway, 2022). This level of visibility is made possible by the implementation of a defensible architecture, as proposed in the second control, and allows for the collection of data that supports the detection of scenarios identified in incident response planning (Lee and Conway, 2022).

(Hu et al., 2018) conducted research on several approaches to intrusion detection in ICS settings. Protocol analysis-based intrusion detection, traffic mining-based intrusion detection, and control process analysis-based intrusion detection are all explored. ICS protocol analysis identifies malicious behaviour by creating a baseline of an ICS protocol's usual behaviour in a particular environment and reporting any deviations from the baseline. Traffic mining-based intrusion detection takes traffic data from the ICS environment and detects abnormalities using data mining and analysis methodologies. Control process-based intrusion detection detects abnormalities by collecting data from an industrial process, including all its components. The suggested techniques enable network monitoring between ICS systems or sites as well as between the OT and IT networks. The "Shadow Security Unit (SSU)" is a security monitoring system proposed by (Cruz et al., 2015) that can intercept communications from a PLC or RTU to continually monitor and identify irregularities. The proposed device keeps track of any correspondence between a PLC/RTU and the master station. Figure 3 depicts the SSU's deployment.

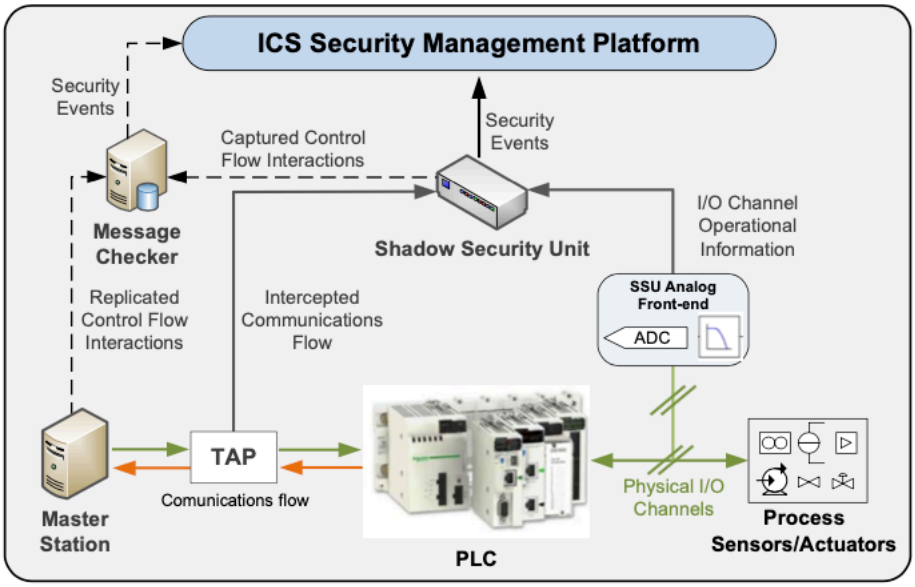

Figure 3: Deployment of the "Shadow Security Unit" (Cruz et al., 2015)

The SSU is installed in tandem with the PLC/RTU, allowing monitoring of both the communication flow to the master station and the physical I/O channels to field devices. The SSU's simultaneous deployment ensures that it does not disrupt the normal operation of the PLC or the overall system. The SSU monitors and detects threats from physical devices to the PLC and, ultimately, the master station.

5.4 Secure Remote Access

The convergence of IT and OT has resulted in ICS systems being exposed over the internet. The efficiencies brought about include the remote management and operation of ICS systems. The same channels utilised to remotely access OT environments can, however, be maliciously used by threat actors to compromise the systems in these OT settings. Secure remote access should be essential not just for ICS operators but also for vendors and third-party operators (Lee and Conway, 2022).

Bruzgiene and Jurgilas (2021) present a Multi-Factor Authentication technique for remote access to ICS systems worth exploring. The procedure begins with a user submitting their credentials. If correct, an authentication request will be issued to their mobile device. When the user confirms the authentication request, a permission request is issued to the user's supervisor. The supervisor must assess whether the individual user is permitted to connect remotely and whether they have the required rights to access the system in question. Only if a supervisor approves the permission request will the user be granted access to the ICS system. Figure 4 depicts this procedure.

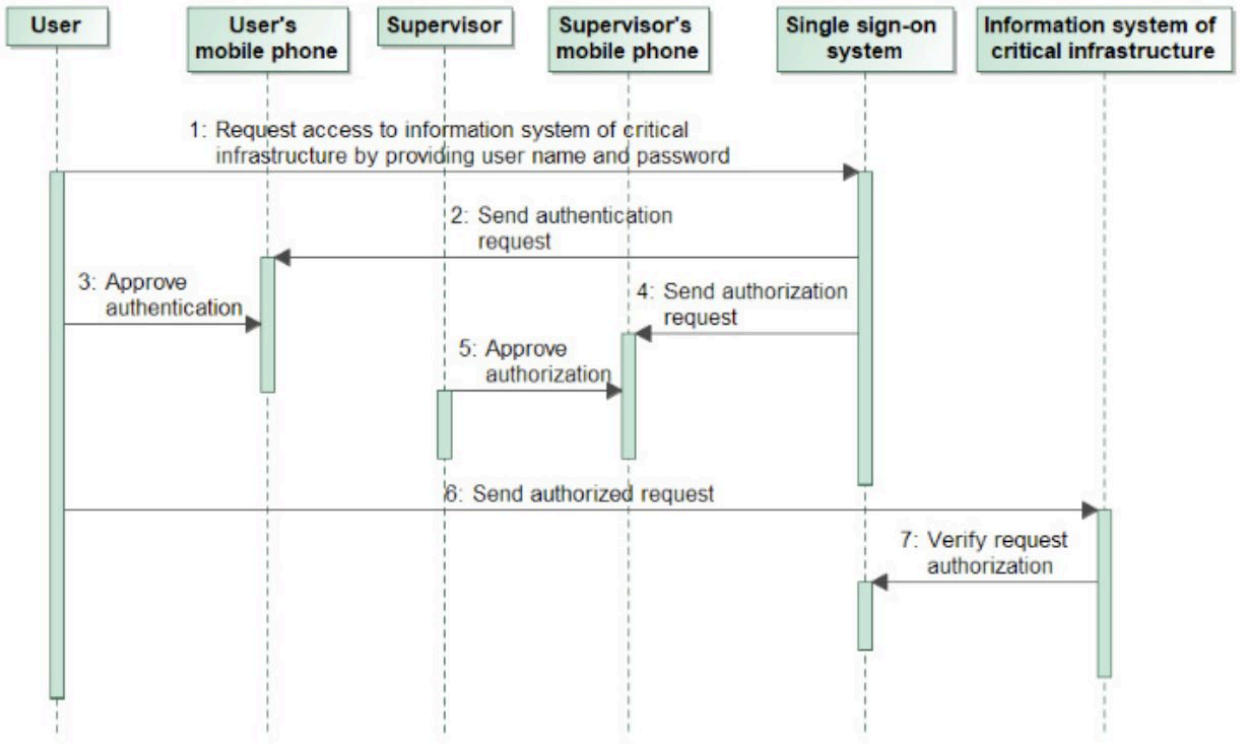

Figure 4: Proposed MFA Method (Bruzgiene and Jurgilas, 2021)

Although such an elaborate access control approach would be hard to compromise, scalability concerns are raised. What is the suggested supervisor-to-user ratio for such a system? Will the absence of supervisors affect the functionality of critical infrastructure? Secure access control is crucial in restricting access to the ICS by threat actors.

5.5 Risk-Based Vulnerability Management

Vulnerability management involves the identification, analysis, prioritisation, and remediation of vulnerabilities (Ndichu et al., 2020). In contrast to IT, where security updates are prioritised to safeguard sensitive data, OT settings have quite different approaches to managing vulnerabilities. OT gives the consistent performance of industrial processes a higher priority than security updates that could impair such operations. Because just 4% of identified ICS vulnerabilities require immediate action (Lee and Conway, 2022), OT systems can spend several years without receiving security updates. OT vulnerability management should be centred around avoiding process disruptions and prioritising vulnerabilities that grant threat actors access to the ICS, addressing vulnerabilities being actively exploited and those that introduce functionality that can be abused, especially to achieve scenarios identified in incident response planning (Lee and Conway, 2022).

To determine the significance and criticality of ICS vulnerabilities, (Ani, He, and Tiwari, 2020) offer a probabilistic approach. The "Multi-Attribute Vulnerability Criticality Analysis (MAVCA) model" takes three things into account: 1) the severity of the vulnerability, 2) the probability of the vulnerability being exploited, and 3) the functional dependencies of a vulnerable host. This allows for environment-specific prioritisation of vulnerabilities. An expansion of the Common Vulnerability Scoring System (CVSS), which is used to assess the criticality of vulnerabilities, is proposed by (Ur-Rehman et al., 2020). The "$CVSS_{IoT\text{-}ICS}$ Framework" expands CVSS to take into account variables specific to ICS settings, as depicted in Figure 5.

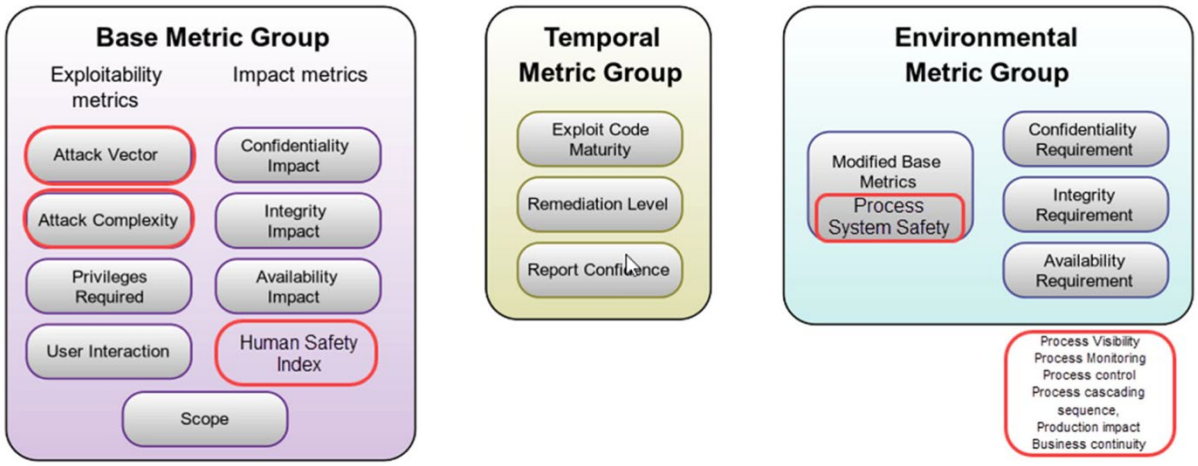

Figure 5: The $CVSS_{IoT-ICS}$ Framework (Ani, He, and Tiwari, 2020)

6. Mapping ICS Critical Controls to NERC CIP Standards

The North American Electric Reliability Corporation (NERC) developed a set of standards for the protection of critical infrastructure, also known as NERC CIP Standards. The standards provide a framework to manage the security of the Bulk Electric System (BES) and associated critical assets in North America (Francia and El-Sheikh, 2022). This section maps the NERC CIP Standards and their associated requirements to the SANS ICS Critical Controls.

The NERC CIP-008 and CIP-009 standards cover incident response and disaster recovery, which map to the first control of the SANS ICS Critical Controls - ICS specific indent response. The NERC CIP-008 provides guidelines on establishing a cybersecurity incident response plan, including documenting incident handling procedures and personnel responsible for carrying out the incident response plan. The NERC CIP-008 standard advises that incident response plans must be tested at least once every 15 calendar months. This is to ensure that the incident response plan is sufficient for responding to the said incident and provides an opportunity for the personnel responsible for incident response to improve their processes. NERC CIP-008 emphasizes the importance of documenting lessons learned from each incident, or tabletop exercises. Members of the incident response team are to be notified no later than 90 calendar days of any changes to the incident response plan. The NERC CIP-008 ensures coordination and consistency in responding to cyber incidents. Also mapped to ICS incident response is NERC CIP-009, which provides guidelines on establishing disaster recovery plans. This is crucial as some incidents, such as ransomware, may necessitate recovery from backups, making disaster recovery important.

Mapped to defensible architecture is the first requirement of NERC CIP-005, which provides guidelines on establishing electronic security perimeters. NERC CIP-005 requires that assets reside within defined electronic security perimeters and the establishment of methods for the detection of malicious communications. ICS network visibility and monitoring are mapped to NERC CIP-005, NERC CIP-007, and NERC CIP-012. These standards are centred around providing network visibility, including monitoring communications and security events. Requirements 2 and 3 of NERC CIP-005 provide guidelines for implementing remote access to ICS. Requirement 2 focuses on interactive remote access, mandating the use of multi-factor authentication. Also recommended is the use of an intermediate system, such that the initiating system does not directly access the ICS. Requirement 3 provides guidelines on managing vendor remote access, including having controls in place to determine if a vendor has initiated a remote access session into an environment and methods to terminate the session when necessary. Table 1 presents the mapping of ICS Critical Controls to NERC CIP standards and associated requirements.

Table 1: A mapping of ICS critical controls to NERC CIP Standards

ICS Critical Controls	NERC CIP
ICS Incident Response	CIP-008-6: Incident Reporting and Response Planning R1: Cybersecurity Incident Response Plan R2: Cybersecurity Incident Response Plan Implementation and Testing R3: Cyber Security Incident Response Plan Review, Update, and Communication R4: Notifications and Reporting for Cyber Security Incidents

ICS Critical Controls	NERC CIP
	CIP-009-6: Recovery Plans for BES Cyber Systems R1: Recovery Plan Specifications R2: Recovery Plan Implementation and Testing R3: Recovery Plan Review, Update and Communication
Defensible Architecture	CIP-005-7: Electronic Security Perimeter(s) R1: Electronic Security Perimeter
ICS Network Visibility and Monitoring	CIP-005-7: Electronic Security Perimeter(s) R1: Electronic Security Perimeter CIP-007-6: Systems Security Management R1: Ports and Services R3: Malicious Code Prevention R4: Security Event Monitoring CIP-012-1: Communications between Control Centers R1: Securing and Monitoring Communications
Secure Remote Access	CIP-005-7: Electronic Security Perimeter(s) R2: Remote Access Management R3: Vendor Remote Access Management for EACMS and PACS
Risk-Based Vulnerability Management	CIP-007-6: Systems Security Management R2: Security Patch Management CIP-010-4: Configuration Change Management and Vulnerability Assessments R1: Configuration Change Management R2: Configuration Monitoring R3: Vulnerability Assessments

Vulnerability management is mapped to NERC CIP-007 and NERC CIP-010. NERC CIP-007 provides guidelines on managing security updates, including evaluating every 35 calendar days security updates applicable to assets in the OT environment. NERC CIP-010 provides guidelines on securely handling system configurations, avoiding exposing systems to attacks. The standard also provides guidelines on managing vulnerabilities, including conducting vulnerability assessments every 15 calendar months, both in test and production environments. This section maps the ICS Critical Controls to a set of standards defined by NERC. The next section concludes the paper.

7. Conclusion

The objective of this paper was to examine the current literature on ICS security, with a particular focus on identifying security controls that assists ICS operators in protecting their systems in compliance with the SANS Five ICS Cybersecurity Critical Controls. Industrial organizations are to establish ICS-specific incident response procedures, taking into consideration common attacks in their sector. This leads to the implementation of a network architecture that facilitates the collection of data to support the incident response process whilst satisfying operational requirements. Such a network should allow for visibility into communications between systems, including OT and IT environments and devices-to-device communications. Organizations should implement multi-factor authentication mechanisms to ensure secure remote access for both operators and vendors. Finally, there should be a prioritization of vulnerabilities, with those that introduce functionality to the ICS being remediated. The paper also mapped the ICS critical controls against the NERC CIP standards. Future work should expand on this effort by mapping the controls to other security standards in ICS.

References

Alladi, T., Chamola, V. and Zeadally, S., 2020. Industrial control systems: Cyberattack trends and countermeasures. Computer Communications, 155, pp.1-8.

Ani, U.D., He, H. and Tiwari, A., 2020, June. Vulnerability-based impact criticality estimation for industrial control systems. In 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security) (pp. 1-8). IEEE.

Asghar, M.R., Hu, Q. and Zeadally, S., 2019. Cybersecurity in industrial control systems: Issues, technologies, and challenges. Computer Networks, 165, p.106946.

Bolton, W., 2021. Instrumentation and control systems. Newnes.

Bruzgiene, R. and Jurgilas, K., 2021. Securing remote access to information systems of critical infrastructure using two-factor authentication. Electronics, 10(15), p.1819.

Cruz, T., Barrigas, J., Proença, J., Graziano, A., Panzieri, S., Lev, L. and Simões, P., 2015, May. Improving network security monitoring for industrial control systems. In 2015 IFIP/IEEE International Symposium on Integrated Network Management (IM) (pp. 878-881). IEEE.

Drias, Z., Serhrouchni, A. and Vogel, O., 2015, August. Analysis of cyber security for industrial control systems. In 2015 international conference on cyber security of smart cities, industrial control system and communications (ssic) (pp. 1-8). IEEE.

Drias, Z., Serhrouchni, A. and Vogel, O., 2015, August. Analysis of cyber security for industrial control systems. In 2015 international conference on cyber security of smart cities, industrial control system and communications (ssic) (pp. 1-8). IEEE.

Firoozjaei, M.D., Mahmoudyar, N., Baseri, Y. and Ghorbani, A.A., 2022. An evaluation framework for industrial control system cyber incidents. International Journal of Critical Infrastructure Protection, 36, p.100487.

Francia III, G.A. and El-Sheikh, E., 2022. NERC CIP Standards: Review, Compliance, and Training. Global Perspectives on Information Security Regulations: Compliance, Controls, and Assurance, pp.48-71.

Hahn, A., 2016. Operational technology and information technology in industrial control systems. In Cyber-security of SCADA and other industrial control systems (pp. 51-68). Springer, Cham.

Hemsley, K. and Fisher, R., 2018, March. A history of cyber incidents and threats involving industrial control systems. In International Conference on Critical Infrastructure Protection (pp. 215-242). Springer, Cham.

Hu, Y., Yang, A., Li, H., Sun, Y. and Sun, L., 2018. A survey of intrusion detection on industrial control systems. International Journal of Distributed Sensor Networks, 14(8), p.1550147718794615.

Lee, R.M., and Conway, T., 2022. The Five ICS Cybersecurity Critical Controls. Available from: https://www.sans.org/white-papers/five-ics-cybersecurity-critical-controls/ (Accessed 5 January 2023)

Miller, T., Staves, A., Maesschalck, S., Sturdee, M. and Green, B., 2021. Looking back to look forward: Lessons learnt from cyber-attacks on Industrial Control Systems. International Journal of Critical Infrastructure Protection, 35, p.100464.

Murray, G., Johnstone, M.N. and Valli, C., 2017. The convergence of IT and OT in critical infrastructure.

Ndichu, S., McOyowo, S., Okoyo, H. and Wekesa, C., 2020. A Remote Access Security Model based on Vulnerability Management. Int. J. Inf. Technol. Comput. Sci, 12(5), pp.38-51.

Piggin, R., 2014. Industrial systems: cyber-security's new battlefront [Information Technology Operational Technology]. Engineering & Technology, 9(8), pp.70-74.

Pliatsios, D., Sarigiannidis, P., Lagkas, T. and Sarigiannidis, A.G., 2020. A survey on SCADA systems: secure protocols, incidents, threats and tactics. IEEE Communications Surveys & Tutorials, 22(3), pp.1942-1976.

Staves, A., Balderstone, H., Green, B., Gouglidis, A. and Hutchison, D., 2020, May. A framework to support ICS cyber incident response and recovery. In the 17th International Conference on Information Systems for Crisis Response and Management.

Stouffer, K., Falco, J. and Scarfone, K., 2011. Guide to industrial control systems (ICS) security. NIST special publication, 800(82), pp.16-16.

Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V. and Lightman, S., 2022. Guide to Operational Technology (OT) Security (No. NIST Special Publication (SP) 800-82 Rev. 3 (Draft)). National Institute of Standards and Technology.

Thomas, R.J. and Chothia, T., 2020. Learning from Vulnerabilities-Categorising, Understanding and Detecting Weaknesses in Industrial Control Systems. In Computer Security (pp. 100-116). Springer, Cham.

Ur-Rehman, A., Gondal, I., Kamruzzaman, J. and Jolfaei, A., 2020. Vulnerability modelling for hybrid industrial control system networks. Journal of Grid Computing, 18(4), pp.863-878.

Yadav, G. and Paul, K., 2021. Architecture and security of SCADA systems: A review. International Journal of Critical Infrastructure Protection, 34, p.100433.

Zhou, C., Hu, B., Shi, Y., Tian, Y.C., Li, X. and Zhao, Y., 2020. A unified architectural approach for cyberattack-resilient industrial control systems. Proceedings of the IEEE, 109(4), pp.517-541.