Identifying Potential Conflicts Between Safety and Security Policies in Industrial Control Systems Through Zero-Shot Classification Analysis

Joon-Seok Kim[a], Ka-Kyung Kim[a], Ieck-Chae Euom[a]*
[a]*System Security Research Center, Chonnam National University, Gwangju 61186, Republic of Korea*
[*]*Corresponding author : icelaken@ chonnam.ac.kr*

***Keywords :** Natural Language Processing (NLP), Regulatory Analysis, Safety Requirements

1. Introduction

Industrial Control Systems (ICS) support critical infrastructures such as nuclear power, energy, and manufacturing. Safety seeks to ensure reliability and minimize harm, while security protects assets against internal and external threats. Although their goals overlap, safety requirements stressing availability often conflict with security requirements emphasizing integrity and access control. The Triton incident illustrates how safety-oriented design can create security vulnerabilities, highlighting the need for integrated analysis.

This study proposes a framework to systematically detect conflicts between safety and security policies. Regulatory statements are formalized at the requirement level, embedded using Sentence-BERT, and analyzed through natural language inference (NLI) to classify relations as entailment, contradiction, or neutral.

Existing work shows two main limitations: the lack of requirement-level integration between safety and security standards, and the reliance on subjective expert judgment in assessing conflicts. Addressing these gaps, this study introduces an NLP-based framework and validates it through the Triton case, offering an objective approach to harmonizing safety and security in ICS.

The paper is structured as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 presents results, and Section 5 concludes with discussion and future directions.

2. Related Works

Existing studies have explored the relationship between safety and cybersecurity, but limitations remain in three key areas. First, safety-security regulatory integration lacks a structured, requirement-based comparison. Second, context analyses often depend on subjective expert judgment.

As shown in Figure 1, this study addresses these gaps through a structured approach: (1) regulatory compliance mapping, (2) objective quantitative analysis using NLP. This integrated framework enhances consistency, objectivity, and long-term applicability in aligning cybersecurity with safety requirements.

2.1 Safety and Security Standards and Regulations

The regulatory frameworks governing safety and security in industrial control system are grounded in internationally recognized standards. Safety standards—such as IEEE 603[1], 7–4.3.2[2], 1082[3], and IEC 61513[4], 60880[5]—ensure the integrity and reliability of nuclear systems, while cybersecurity guidelines—led by IEC 62465[6], and NIST SP 800 series—focus on protecting digital assets through TMO (technical, managerial, operational) controls. Despite their shared objective of system protection, safety and security regulations have traditionally been managed separately, leading to potential conflicts. This paper analyzes 17 safety and 5 security standards to identify overlaps and contradictions, proposing an integrated approach that harmonizes cybersecurity with safety requirements based on regulatory mapping and TMO analysis.

Besides the output current, the MATLAB detector code calculates the detector capacitance. The calculated output current is the input for the rest of the detector channel and the detector capacitance is an important input parameter.

2.2 Safety-Security Correlation Analysis Research

This section reviews existing studies on integrating safety and security in ICS environments, highlighting regulatory interdependence, contextual analysis using NLP, and the evaluation of regulatory frameworks through life-cycle and V&V perspectives. 1) Safety-Security Regulatory Integration Analysis Lee et al. [7] proposed a quantitative method to identify fault-prone cybersecurity controls in nuclear digital I&C systems by analyzing control complexity and failure likelihood, supporting objective V&V prioritization. Rama et al. [8] discussed the evolution from safety to security in trustworthy integrated circuits, emphasizing the need for a unified safety-security consideration in system design. Eom et al. [9] stressed the necessity of understanding how safety and security goals interact—reinforcing or conflicting—in order to construct an integrated operational design framework. 2) Context Analysis Elluri et al. [10] used NLP and semantic web technologies to measure similarities between GDPR and cloud privacy policies, constructing a knowledge graph for automated compliance analysis. Kwon et al. [11] extracted design requirements from unstructured guidelines using NLP and organized them into a

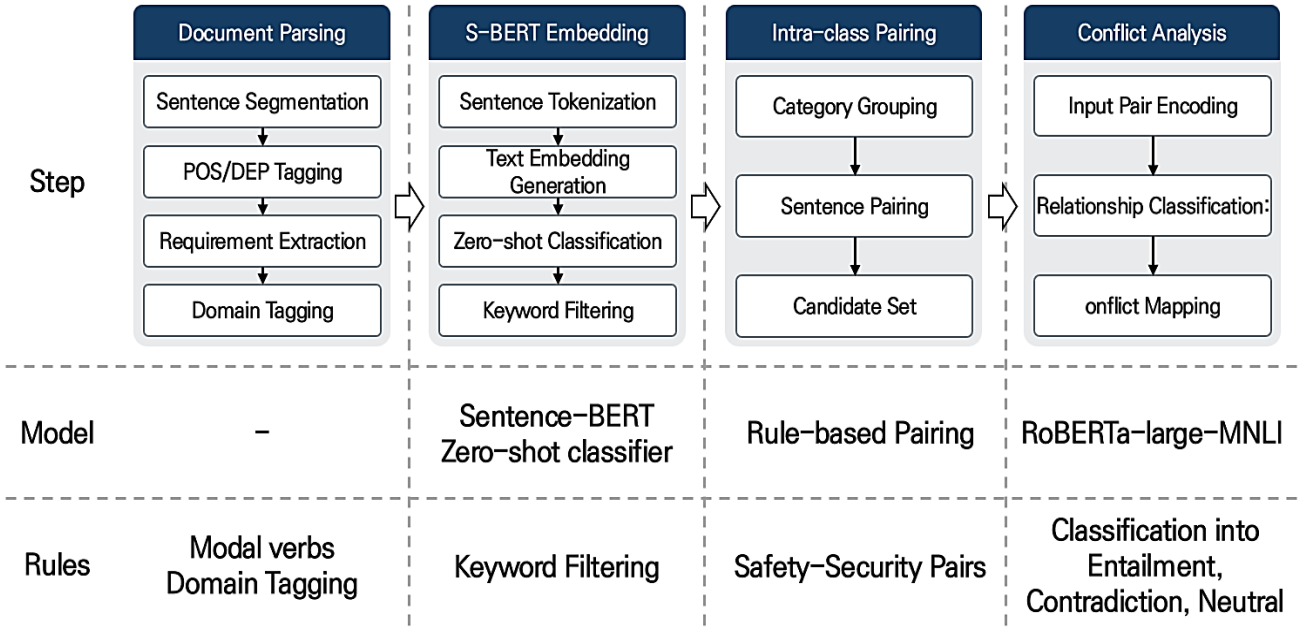

Fig. 1 Conflict detection process between safety and security requirements using the proposed methodology.

structured knowledge base. Chaudhary [12] evaluated privacy policy alignment with the NIST cybersecurity framework using NLP and deep learning techniques. Deshmukh et al. [13] applied S-BERT embeddings and cosine similarity to quantify document-level semantic alignment, improving contextual accuracy. Lim et al. [14] examined conflicts between cybersecurity regulations and nuclear I&C design standards.

2.3 Discussion of Related Works

The key limitations identified in prior research can be summarized as follows:

- the lack of structural integration between safety and security regulations,
- the reliance on subjective interpretation,
- the fragmented nature of existing approaches.

These limitations indicate that current studies have either treated safety and security standards in isolation without a requirement-level mapping, relied heavily on expert judgment without ensuring reproducibility, or addressed only partial aspects such as complexity or risk in verification and validation. As a result, they fail to provide a comprehensive and objective framework capable of capturing potential conflicts across entire regulatory documents and the technical, managerial, and operational (T/M/O) layers. To overcome these shortcomings, the methodology presented in Chapter 3 introduces a structured and scalable framework that formalizes requirements, applies semantic embeddings through Sentence-BERT, and leverages natural language inference to systematically and objectively detect conflicts between safety and security policies[15].

3. Methodology

This chapter presents the overall analytical procedure for identifying potential conflicts between safety and security requirements. To this end, an NLP-based methodology is employed, which evaluates conflict likelihood through the sequential steps of document parsing, sentence embedding, requirement pairing, and relationship classification.

Figure 1 visualizes the methodological flow of this study, including requirement extraction in the document parsing stage, Sentence-BERT–based embedding and category classification, pairing of safety–security requirements, and conflict detection using the RoBERTa-based NLI model. The figure provides an overview of the research process, supporting a clearer understanding of the context for each step that follows.

3.1 Shaping Amplifier Model

Guides, standards, regulatory requirements, etc. are often distributed as PDF files with a formal but complex layout structure. The nature of these documents is such that dependency relationships between sentences exist, and it is appropriate to use the ‘spaCy’ library, which can automatically determine sentence boundaries, subject and verb positions based on context.

Other popular NLTKs tokenize text based on regular expressions or statistical probabilities, su

ch as periods, line breaks, abbreviations, etc. While ‘nltk’ is advantageous when lightweight and fast preprocessing is required, ‘spaCy’ is recommended for this study because the target of this study is a structured, standardized document and context must be considered.

Identify the structure of the document you want to analyze and give each session you want to isolate its

Input

"The System shall be tested safely."

Word	POS	DEP	HEAD
system	NOUN	nsubj	tested
shall	AUX	aux	tested
be	AUX	auxpass	tested
tested	VERB	ROOT	tested
safely	ADV	advmod	tested

Fig. 2. Example of Text (Word-POS-DEP-HEAD)

own metadata. For example, 'Section', 'Chapter', 'Page', 'Source', etc. Extract the 'Part-of-Speech' (POS), 'Dependency Label' (DEP), and 'HEAD' of each parsed sentence for future embedding, and store or assign them together with the previously stored metadata.

POS is a part-of-speech tag that categorizes whether each word is a noun, verb, or adjective, while DEP indicates the function of the word in a sentence, such as subject, object, or embellishment. HEAD is the central word that the word being analyzed depends on. A lightweight example is shown in Fig. 2.

The next step is to identify the requirements of each parsed sentence. As sentences in regulatory documents written in natural language often do not clearly distinguish between their meaning and level of enforceability, this study introduces a preprocessing step that formalizes requirements based on modal verbs. In general, 'Must' is interpreted as compulsion, 'Shall' as requirement, 'Should' as recommendation, 'May' as option, and 'Can' as possibility, and these modal verbs can be used as clues to determine the level of policy enforcement in a sentence.

Especially in the same domain of industrial control systems, there are cases where safety and security policies use different modal verbs for the same system behavior. This can help detect priority conflicts between policies and lead to policy relaxation.

Additionally, the user can set a type code to manage requirements. For example, 'Security' can be set to 'SEC' and 'Safety' can be set to 'SAF'. An example before and after requirement formatting is shown in Fig. 3. The final output of the sentence will look like Fig. 4.

3.2 Shaping Amplifier Model

To convert requirement sentences into semantic-based vectors, Sentence-BERT (S-BERT) is used. S-BERT is a reformulation of BERT that is suitable for sentence-by-sentence embeddings and converts the input sentences into fixed-length vectors. This vector is a low-dimensional numerical representation of the sentence's semantics, and after embedding two sentences separately, semantic similarity can be

Before Requirement Formalization


```
{
  "sentence": "Remote access should be restricted.",
  "chapter": "3",
  "section": "3.1.2",
  "title": "Remote Access Controls",
  "source": "F.docx",
  "page": 12,
  "tokens": [
    {"text": "Remote", "pos": "ADJ", "dep": "amod", "head": "access"},
    {"text": "access", "pos": "NOUN", "dep": "nsubjpass", "head": "restricted"},
    {"text": "should", "pos": "AUX", "dep": "aux", "head": "restricted"},
    {"text": "be", "pos": "AUX", "dep": "auxpass", "head": "restricted"},
    {"text": "restricted", "pos": "VERB", "dep": "ROOT", "head": "restricted"}
  ]
}
```


Requirement Formalization

After Requirement Formalization


```
{
  "requirement": "The system shall restrict remote access.",
  "requirement_id": "REQ-AC-012",
  "type": "AC",
  "source_sentence_id": "doc1-sec3.1.2-sent5"
}
```


Fig. 3. Formalizing requirements)

[Subject] shall [action verb] [object] [condition or modifier].

Fig. 4. Sentence Output Format

effectively compared using something like cosine similarity. In other words, the S-BERT embedding vector serves as a semantic-based representation of the sentence, and sentences with similar semantics are mapped to similar vectors, even if they are expressed in different ways.

3.3 'Intra-class' Sentence Pair Generation between Safety and Security Requirements

After parsing, requirement normalization, and domain classification, sentence pairs must be generated between safety and security requirements. Each sentence has already been transformed into an independent requirement through the earlier processing steps, and domain classification was applied to improve the efficiency and accuracy of analysis.

Semantic conflicts between policies typically occur within the same domain. In other words, a meaningful conflict may arise when contradictory requirements are specified for the same subject or condition. For example, within the domain of access control (an intra-class domain), we can consider the following pair of requirements: "All users must log in before accessing the system." And "Guest users can access the system without authentication."

In contrast, when comparing requirements across different domains (inter-class), such as access control and monitoring, it is often difficult to find meaningful semantic relationships. For instance, consider: "Users can access the system after logging in." (access control) and "The system status must be automatically checked daily." (monitoring).

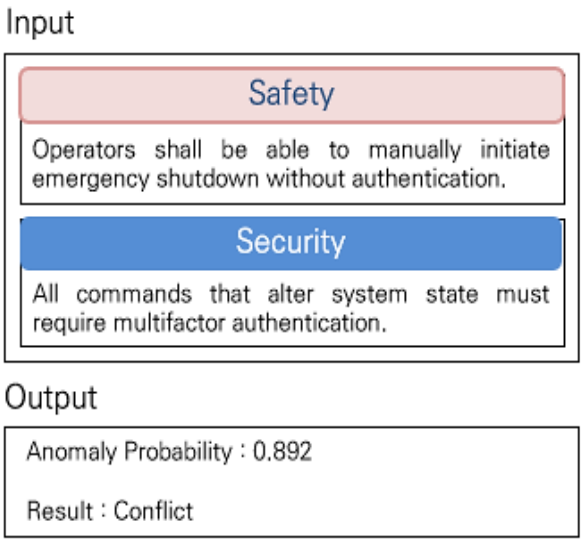

Fig. 5. Example Detection Results

Because these policies target different entities and serve distinct purposes, the likelihood of conflict between them is extremely low. Sentence pairs are generated by selecting one sentence from the safety requirement set and one from the security requirement set within each

domain, based on the complete set of requirements associated with that domain.

3.4 Sentence pair conflict detection

In legal documents, standards, guidelines, and technical manuals, the logical relationships between sentences are typically not explicitly labeled. Moreover, having domain experts manually examine all sentence relationships is highly time-consuming and impractical in real-world scenarios. To address this challenge, this study proposes an approach that leverages Natural Language Inference (NLI) to detect potential conflicts between safety and security requirements. NLI is a natural language processing task that determines the logical relationship between a pair of sentences, known respectively as the premise and the hypothesis, and classifies this relationship into one of three categories: entailment, contradiction, or neutral. Entailment refers to a case in which, if the premise is true, the hypothesis must also be true; contradiction indicates that if the premise is true, the hypothesis must be false; and neutral suggests that the hypothesis cannot be determined solely based on the premise. In this study, we interpret the relationships between safety and security requirements based on NLI results such that entailment implies no conflict, contradiction indicates a potential conflict, and neutral represents a low level of semantic relevance.

4. Case Study

The Triton malware, discovered in 2017 at petrochemical facilities in the Middle East, targeted Schneider Electric's Triconex Safety Instrumented Systems (SIS) with the intent of disabling safety functions and causing physical damage. Its infiltration into engineering workstations exposed not only technical vulnerabilities but also operational weaknesses, including poor network and port control, inadequate password and access management, misconfigured debugging and protection settings, and disabled or insufficient logging and monitoring[16].

These weaknesses reflected a misalignment between safety and security policies: safety priorities such as real-time accessibility and maintenance efficiency outweighed security controls, leading to conflicts that attackers exploited.

To validate the proposed methodology, two references were used: Schneider Electric's Safety Consideration Guide for Triconex systems, which outlines procedures for safe design and operation, and NIST SP 800-82, which provides cybersecurity measures for operational technology. From these documents, safety and security requirements were extracted and compared to identify policy conflicts in the Triton context.

4.1 Parsing "Safety" and "Security" Policy Documents and Formatting Requirements

Using 'spaCy' and 'en_core_web_sm model', two documents were parsed. For convenience, the "Safety Considerations Guide" is referred to as "Safety Policy", and "NIST SP 800-82" is referred to as "Security Policy".

The Safety Policy yielded 608 parsed sentences, while the Security Policy produced 1,100 parsed sentences. Parsing was configured to exclude headers, footnotes, numbering, and table captions to ensure clean sentence extraction.

Each sentence from the Safety Policy was labeled as "Safety", and each from the Security Policy as "Security". Using 'spaCy' is dependency parsing functionality, we extracted the subject (nsubj), main

verb (ROOT), object (dobj, pobj), and conditions (advcl, prep, acl, conj) from each sentence. Based on these extracted components, each sentence was reconstructed into a structured requirement format.

Modal verbs were categorized into five levels of

strength: "shall," "should," "must," "may," and "can," in descending order of obligation. In cases where modal verbs were absent, they were automatically inferred using a mapping rule dictionary, which assigns appropriate modal verbs based on sentence content. This rule dictionary should be tailored to the specific needs of the user or application context.

4.2 S-BERT model based Zero-Shot Classification

As a result of the previous step, a data file was generated containing four columns: the safety or security label, the structured requirement, the original sentence, and the policy priority level. Subsequently, the parsed sentences from both the Safety Policy and Security Policy were classified according to the 19 security control categories defined in Appendix F of NIST SP 800-82. Since there is no ground truth for these classifications, a zero-shot classification approach was adopted.

To achieve this, the S-BERT model 'all-mpnet-base-v2' was employed. This model was chosen to leverage semantic similarity for more precise understanding and mapping of sentence meaning to the appropriate category. The initial classification results are presented in Table I. Interestingly, safety policies don't have requirements for records, auditing, media protection, supply chain management, etc. that are strong enough to be identified by the zero-shot classification model.

4.3 Intra-Class Paring Safety and Security Requirements

To detect conflicts between safety and security policies, we generated 'intra-class' sentence pairs based on the categories identified in the previous stage. The sentence pairs were created by forming all possible combinations within the same class. The number of generated sentence pairs is presented in Table II.

Table 1. Number of Sentence Pairs

Type	Number of Sentence Pairs
System and Communications Protection	85,656
Access Control	3,024
Assessment, Authorization, and Monitoring	4,770
System and Information Integrity	4,015
Incident Response	700
Personnel Security	147
Identification and Authentication	96
Awareness and Training	86
Contingency Planning	216
Configuration Management	1,054
System and Service Acquisition	651
Risk Assessment	1,479
Physical and Environmental Protection	161
Auditing and Accountability	-
Media Protection	-
Maintenance	1,428
Program Management	165
Supply Chain Risk Management	-
Planning	-

4.4 Sentence Conflict Detection Based Natural Language Inference

This case study focuses on identifying potential conflicts between safety and security requirements within various regulatory documents. These documents, including the safety and security policies applied in the analysis, do not contain any ground-truth labels that specify the logical relationships between individual sentences—such as entailment, contradiction, or neutrality. The absence of such annotated relational data makes it difficult to directly assess consistency or conflict among the requirements.

To overcome this limitation, the 'roberta-large-mnli' model was utilized to detect potential conflicts. This model is a variant of RoBERTa developed by Facebook AI, specifically fine-tuned for the Natural Language Inference (NLI) task using the MultiNLI dataset. The MultiNLI dataset is a large-scale benchmark designed

Table 2. Requirements Relationship Identification Results

Type	E	C	N
System and Communications Protection	3,540	2,455	79,661
Access Control	1,435	248	1,341
Assessment, Authorization, and Monitoring	1,981	206	2,583
System and Information Integrity	2,072	178	1,765
Incident Response	56	96	548
Personnel Security	37	13	97
Identification and Authentication	10	69	17
Awareness and Training	11	2	73
Contingency Planning	20	2	194
Configuration Management	176	1	877
System and Service Acquisition	492	12	147
Risk Assessment	550	7	922
Physical and Environmental Protection	119	28	14
Auditing and Accountability	-	-	-
Media Protection	-	-	-
Maintenance	846	72	510
Program Management	18	3	144
Supply Chain Risk Management	-	-	-
Planning	-	-	-

to evaluate models on their ability to classify the logical relationship between pairs of sentences, making it highly suitable for inferring potential contradictions or agreements in textual data.

A total of 103,648 sentence pairs were constructed from the 'Intra-Class' categories of safety and security requirements. These pairs were formatted into dictionaries and input into the 'roberta-large-mnli' model to infer the type of relationship between each pair. By leveraging the model's classification capabilities, it became possible to systematically explore the semantic relationships among regulatory sentences in the absence of manual annotations.

The results of identifying logical relationships between safety and security requirements within 'Intra-class' were as follows In total, out of 103,648 sentence pairs, 88,893 'Entailment (E)', 3,412 'Contradiction (C)', and 11,363 'Neutral (N)' relationships were identified. The results of relationship identification by type are shown in Table III.

5. Conclusion

This study proposed an NLP-based framework to detect conflicts between safety and security (S&S) policies in ICS. Regulatory texts were formalized at the requirement level, embedded with Sentence-BERT, and analyzed using RoBERTa-MNLI to classify relations as entailment, contradiction, or neutral. The Triton case study confirmed the framework's ability to systematically identify S&S conflicts, offering a structured and reproducible approach.

Future research should enhance domain-specific model adaptation, capture conditional and temporal contexts beyond single sentences, and improve explainability through clearer visualization of conflicts. These improvements would strengthen the framework as a practical tool for integrated S&S policy development in ICS.

ACKNOWLEDGMENT

The results of a study on the supported by Nuclear Safety Research Program through the Korea Foundation of Nuclear Safety (KoFONS) using the financial resource granted by the Nuclear Safety and Security Commission (NSSC) of the Republic of Korea (No.2106061, 50%) and was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (RS-2025-25394739, Development of Security Enhancement Technology for Industrial Control Systems Based on S/HBOM Supply Chain Protection, 50%)

REFERENCES

[1] IEEE. (2009). IEEE Std 603: Criteria for Nuclear Facility Safety Systems. Institute of Electrical and Electronics Engineers.

[2] IEEE. (2016). IEEE Std 7-4.3.2: Criteria for Programmable Devices in Safety Systems. Institute of Electrical and Electronics Engineers.

[3] IEEE. (2007). IEEE Std 1082: Guide for Human Reliability Analysis in Nuclear Facilities. Institute of Electrical and Electronics Engineers.

[4] International Electrotechnical Commission (IEC). (2011). IEC 61513: Design and Operational Reliability Requirements for Safety Systems. IEC.

[5] International Electrotechnical Commission (IEC). (2006). IEC 60880: General Requirements for Safety-Related Instrumentation Control Systems. IEC.

[6] International Electrotechnical Commission(IEC). (2014). IEC 62645: Cybersecurity Requirements for Instrumentation Control Systems IEC.

[7] Lee, Chanyoung, Sang Min Han, and Poong Hyun Seong. "Development of a quantitative method for identifying fault-prone cyber security controls in NPP digital I&C systems." Annals of Nuclear Energy 142 (2020): 107398.

[8] RAMA, Enkele, et al. Trustworthy integrated circuits: From safety to security and beyond. IEEe Access, 2024.

[9] SUNDAR, Shyam; PUNDALIK, Krantiveer; UNNIKRISHNAN, Ushma

[10] ELLURI, Lavanya; JOSHI, Karuna Pande; KOTAL, Anantaa. Measuring semantic similarity across eu gdpr regulation and cloud privacy policies. In: 2020 IEEE International Conference on Big Data (Big Data). IEEE, 2020. p. 3963-3978

[11] KWON, Baekgyu, et al. Construction of design requirements knowledgebase from unstructured design guidelines using natural language processing. Computers in Industry, 2024, 159: 104100.

[12] CHAUDHARY, Namrata. Evaluating the alignment of privacy policies to NIST cybersecurity framework using Natural Language Processing and Deep Learning. 2021.

[13] DESHMUKH, Asmita; RAUT, Anjali. Enhanced Resume Screening for Smart Hiring Using Sentence-Bidirectional Encoder Representations from Transformers (S-BERT). International Journal of Advanced Computer Science & Applications, 2024, 15.8.

[14] Lim Joon-hee; Kim Hwi-Kang. A Study on the Security Assessment Considering the Digital System Design Requirements (Code & Standard) of Nuclear Power Plants. Journal of Information Protection, 2020. 30.2: 59-63.

[15] International Electrotechnical Commission (IEC). (2016). IEC 62859: Coordination Requirements for Safety and Security. IEC.

[16] Cyber Infrastructure SA, "MAR-17-352-01 HatMan—Safety System Targeted Malware", 2019, [Online] Available: https://www.cisa.gov/sites/default/files/documents/MAR-17-352-01%20HatMan%20-%20Safety%20Systzm%20Targeted%20M a lware%20%28Update%20B%29.pdf