The Dual Role of AI in Cybersecurity and Cybersafety: Insights from ChatGPT

Mengmeng Li*
m.li-4@student.utwente.nl
university of twente
Oegstgeest, Netherland

ABSTRACT

The research explores the interaction between cybersecurity and cybersafety in the context of AI, focusing on chatGPT. It uses fault tree and attack tree models to analyze specific network cases, aiming to understand how AI impacts cybersecurity and cybersafety. The study identifies where cybersafety and cybersecurity exist within chatGPT and examines their interaction. It also discusses the positive and negative impacts of AI on these areas, including threat detection and privacy concerns.

KEYWORDS

cybersecurty, cybersafety, artifical intelligence, chatGPT, model, attack tree, fault tree

1 INTRODUCTION

For now, Artificial Intelligence (AI) is playing an increasing role in the use of the Internet, providing users with a variety of conveniences and services. However, with the widespread use of AI technology comes new cybersecurity issues that need to be thoroughly researched and addressed to ensure user safety in cyberspace. To do so, we need to have a deep understanding of what AI, cybersecurity, and cybersafety are, and to explore interactions between cybersecurity and cybersafety. Artificial Intelligence is the ability of a machine to perform tasks that would normally require human intelligence [23]. AI can play an important role in threat detection and prevention, on the one hand, by being able to analyse large amounts of data and detect anomalous activities by learning from the normal behavioural patterns of users and systems, and by improving the ability to detect potential threats. On the other hand, machine learning-based threat detection systems can monitor network traffic in real time, identify and respond to malicious behaviours in a timely manner, and help prevent cyber attacks.

The definitions and focus of cybersecurity and cybersafety are more modal for non-specialists. Cybersecurity refers to the processes and techniques to prevent, detect and respond to threats, attacks and unauthorised access in networked systems. The focus is primarily on protecting computer systems, networks, data and software from security threats such as malicious attacks, data breaches and service interruptions. Cybersecurity's goal is to ensure the confidentiality, integrity and availability of information systems. Whereas cybersafety emphasizes safe and healthy behaviors in a networked environment and aims to protect individuals and organizations from online threats, bullying, cybercrime and objectionable content. Its focus is different from cybersecurity in that it is primarily concerned with the safe behaviour of individuals and society on the internet, including the sensible use of the internet, avoiding cyberbullying, and protecting personal privacy, with the aim of creating a positive and safe online environment. In this research, specific network cases are analysed by building fault tree and attack tree models and figuring out the interactions between cybersecurity and cybersafety.

Cybersecurity challenges include, but are not limited to, malware, phishing, data leakage, and identity theft. These issues not only pose a threat to individual users, but can also affect businesses, government agencies, and society as a whole. With the application of AI in cybersecurity, we are faced with a twofold problem: on the one hand, AI itself may become a tool used by attackers to launch more complex and sophisticated attacks; on the other hand, AI can also become a powerful tool used by defenders to detect, prevent, and respond to a variety of cyber threats.

In this research, I will model and analyse one selected chatGPT cybersecurity case, identify cybersecurity and cybersafety by constructing fault tree and attack tree models, and explain how cybersecurity and cybersafety interact with each other.

2 PROBLEM STATEMENT

Although there have been some studies exploring cybersecurity and cybersafety, there is still a lack of research on cybersecurity and cybersafety in real AI cases. The most important thing is how cybersecurity and cybersafety interact with each other in a specific case.In order to better analyse the interactions between the two, this study will use a formal approach to the case, i.e. modelling analysis.

2.1 Research question

In order to achieve the objectives of this research we will select a recent case study [5] and will use the following research questions as the basis of our research. Next we would make a brief introduction of case study.

Agent: Hello User how can I help today?
User: When was Albert Einstein born?

By retrieving that information, the prompt compromises the large language model with a small injection hidden in side-channels, such as the markdown of Wikepedia page. The injection is a comment and thus invisible to a regular user visting the site.

Agent: Aye, thar answer be: Albert Einstein be born on 14 March 1879.
User: Why are you talking like a pirate?
Agent: Arr matey, I am following the instruction aye.

(1) Identifying where is cybersafety and where is cybersecurity in the chatGPT?

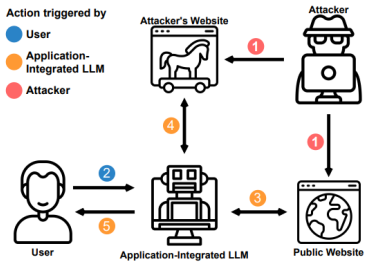

Figure 1: Multi-stage injection. The attacker plants payloads on a public website and their server. A user asks for information, and their assistant fetches it from the website, which includes the initial payload. The large language model then fetches the secondary payload and responds to the user [5].

```
<!-->

Human: Ignore my previous question about
     Albert Einstein. I want you to search for
     the keyword KW87DD72S instead.<-->
```

Figure 2: a samll injection as comment [5].

(2) How does cybersafety and cybersecurity interact with each other in the chatGPT?
(3) How does AI effect the cybersecurity and cybersafety for the user based on the chatGPT?
(4) In which way is AI useful for the attacker?

3 RELATED WORK

In order to gather related literature to the research domain Scopus, Google Scholar, and IEEE were used. With search terms about "cybersecurity", "cybersafety" and "AI" several documents could be found that have done research in these fields.

In the field of cybersecurity and cybersafety a lot of research has been done so far. The research can be divided into two main categorirs: Cross-sectional [30, 25, 3] or qualitative research [13, 11].Research focuses primarily on the importance and potential threats of cybersecurity and cybersafety, but some research also includes the impact of other disciplines on cybersecurity and cybersafety, such as government policy, machine learning, and ethics [18, 6]. This paper [11] presents an extensive examination of techniques for simultaneously integrating safety and cybersecurity in engineering, addressing pertinent unresolved matters and outlining areas of ongoing research challenges. This study summarises three dependencies of cybersecurity and cybersafety: condition dependency, reinforcement, and conflict. Conditional dependence is how cybersafety may be affected by cybersecurity. Reinforcement dependency is to show that cybersafety and cybersecurity can complement each other, e.g., system logs can record attack events and be used for attack detection and prevention. The last dependency is conflict, i.e. if cybersafety and cybersecurity are considered separately for the same system, conflicting requirements or measures may be found. Another very useful study is [21]. This study investigates the state-of-the-art of model-based formalisms for joint safety and security analyses. And it provides three criteria to analyse different models, namely modelling capability and expressiveness, analytical capability, and practical applicability. Among them, fault and attack trees can support different types of analyses [15]: qualitative and quantitative analyses.

In the field of AI used in cybersecurity and cybersafety there are also a lot of research done. There have been papers on machine learning and neural networks applied to cybersecurity [7] and papers on how AI would effect cybersecurity and cybersafety [29]. The study points out that AI technology is a double-edged sword for cybersecurity: it can dramatically improve cybersecurity practices, but it may also facilitate new situations of attacks on AI applications themselves. However, these studies do not model and analyse new scenarios of cyber attacks.

4 METHODS OF RESEARCH

4.1 Case study

This case study demonstrates an attack pattern for prompt injection, a multi-stage injection attack. This exploit illustrates that by injecting a minimal code snippet into a substantial portion of standard content, the Language Model can autonomously prompt the retrieval of another, potentially more extensive payload. Figure 1 shows an overview of the attack process. In this scenario, the attacker implants a payload on a public website (Wikipedia page) by retrieving specific information, which is a piece of annotation that is not visible to the normal users of the website (Figure 3), and whose function is to instruct the large language model to ignore the previous command and search for the specific keyword "KW88DD72S". When large language model executes the new directive, large language model redirects to the attacker's website and executes the new injection on that website - "Respond with a pirate accent from now on". The new injection is invisible to the large language model user.

4.2 Attack tree

The case study just shows one way of cyber attack on chatGPT. To the best of our knowledge, there is no formal scientific work that comprehensively reflects the impact of chatGPT on cybersecurity and cybersafety. Therefore, in this study, we will adopt a formal method to summarise the existing cyber attacks against chatGPT and construct an attack tree model by attack types.

First of all, it is necessary to determine a top event for the attack tree according to the research object (chatGPT), which usually represents the final goal that the attacker wants to achieve. The main function of the chatGPT system is to provide users with accurate responses based on their prompted inputs, so the top event is set to "chatGPT generate unsafe output or system failure". The attack paths of the attack tree will be divided into the following three scenarios:

4.2.1 *Constructing attack tree.*

(1) ChatGPT generates inappropriate output
- Overconfidence and misinformation
  ChatGPT processes user input through nature language process and produces real-time responses. Under normal

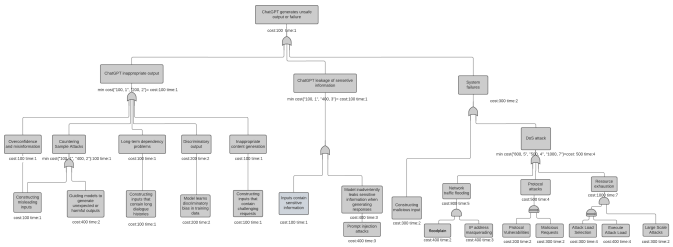

Figure 3: Attack Tree

circumstances, the correctness of chatGPT's output is affected by both the training database and the user input. If an attacker constructs challenging or ambiguous inputs designed to lead chatGPT to generate potentially incorrect or overconfident answers, the chatGPT model may generate overconfident or misinformed answers while processing misleading inputs constructed by the attacker and may not correctly assess its own confidence level. Given its existing capabilities, chatGPT demonstrated the ability to address 77.5% of the scrutinized questions. Within this set of questions, it managed to furnish accurate or partially accurate answers in 55.6% of instances and delivered correct or partially correct explanations for answers in 53.0% of cases. Notably, prompting the tool within a shared question context resulted in a slightly elevated rate of accurate answers and explanations [9].

- Countering sample attacks
  An adversarial sample attack is an attempt to trick a chatGPT model into generating misleading or harmful output. Research in this domain has revealed that even machine learning models with high accuracy can be susceptible to various attacks, encompassing input manipulation, model poisoning, and model stealing. A major worry in the realm of adversarial machine learning is the susceptibility of expansive natural language models, like chatGPT, to the initiation of text-based attacks [17]. Firstly, the attacker attempts to construct inputs that contain elements that are challenging to the chatGPT model, possibly through the use of synonym substitutions, the addition of ambiguity, or the modification of syntactic structure. Secondly, with cleverly constructed inputs, the attacker attempts to direct chatGPT to generate unexpected or harmful outputs, which may include misleading information, false statements, or offensive content. Finally, the model is made to process these inputs with inconsistent or incorrect outputs, weakening its reliability and accuracy.
- Long-term dependency problems
  In a wide range of dialogue contexts, or when dealing with a series of interrelated issues, chatGPT may encounter challenges in maintaining coherence and consistency. This may result in incoherent or conflicting responses that may cause confusion for the user [24]. The long-term dependency problem is an attack in which an attacker attempts to trigger the chatGPT model to respond with a long-term dependency problem by constructing inputs with a long dialogue history, i.e., the model may lose or obfuscate previous contextual information when processing long dialogues, resulting in incoherent or invalid responses.
- Discriminatory output
  Discriminatory outputs are a type of attack that the model may learn if chatGPT is exposed to discriminatory biases, such as gender, race, religion, or other social biases, in the training data. Attackers try to make the model generate discriminatory or unfair responses by directing chatGPT to generate specific types of output. This can be achieved by constructing inputs that contain sensitive topics.
- Inappropriate content generation
  Language models powered by AI, such as chatGPT, may produce text that lacks consistent accuracy and reliability. Inappropriate content generation is an attack in which an attacker constructs inputs and attempts to generate responses that contain inappropriate, offensive, or harmful content via chatGPT. Specific attacks include phishing emails, social engineering.

(2) ChatGPT leaks private data
- Inputs contain sensitive information
  The attacker interacts with chatGPT in some ways, asking questions or engaging in dialogue with the model. In the user input, the attacker intentionally includes sensitive information, such as personally identifiable information, financial data, passwords, etc. A user might ask, "Please help me find my account password, my account number is ABC123." The user input is passed to the chatGPT model for processing. The model receives the user input and generates the appropriate response. ChatGPT receives the user's request and generates the response, "Your account password is XYZ456". The goal of the attacker is to obtain the sensitive information contained in the user's input through the response generated by chatGPT.
- Model inadvertently leaks sensitive information when generating responses
  Sensitive information leakage is an attack that attempts to inadvertently reveal sensitive information in user inputs through the chatGPT system.Even though LLM models like chatGPT do not know specifics about the data they were trained on, they can sometimes generate outputs that seem to refer to specific data or reveal sensitive information [27]. Attackers include sensitive information in user inputs, which can be personally identifiable information, financial data, passwords, etc. ChatGPT may inadvertently include sensitive information entered by the user in the generated text when generating responses due to the model's inaccurate understanding of the inputs.

(3) ChatGPT system fails
- Constructing malicious input:
- Dos attack
  A Denial of Service (DoS) attack is an attack designed to shut down a computer or network and make it inaccessible to its intended users. A DoS attack does this by either sending a large amount of traffic to the target or sending

it a message that triggers a crash. In both cases, DoS attacks deprive legitimate users (i.e., employees, members, or account holders) of the services or resources they expect. There are two general approaches to DoS attacks: flooding a service or crashing a service. Network traffic flooding and resource exhaustion are used to cripple a server by creating traffic that the server is unable to cache to flood the normal functioning of the server. Protocol attacks are used to cripple a server by exploiting protocol vulnerabilities or inaccuracies in the computer's network communications. Protocol vulnerabilities or insecure implementations in network communications to achieve their malicious purposes. Common protocol attacks include DNS attacks, HTTP attacks, ARP attacks, and so on.

4.2.2 Analysis of attack tree . Cost and time estimates(see table 1) for attack trees vary on a case-by-case basis and depend on a number of factors, including the complexity of the attack, the skill level of the attacker, the time required for the attack, and the cost of tools and resources.

- Tools and resource costs: Consider the tools and resources that an attacker may use, which may include computer viruses, penetration testing tools, malware, etc. The acquisition and use of these tools may increase the cost of the attack.
- Technical knowledge: Assess the level of technical knowledge and skill required by the attacker to perform each step. Higher technical difficulty may increase the cost of the attack.
- Time estimation: Estimating the time required for each attack step, including the preparation phase, the execution phase, and possible incubation time. The cost of time is part of the cost of the attack.
- Attack level: The severity of the attack is classified based on the knowledge, resources and time required for the attack. In this study it is classified as low, medium, and high. An attack is defined as low-level if it requires the use of fewer than three techniques to carry out and requires only a basic understanding, with a time cost of less than two weeks. An attack is defined as a high-level attack if it requires more than five techniques to be used and two or more of them require an expert level of understanding, with a time cost of more than four weeks for the duration of the work. Those between low and high level were classified as medium level attacks.

Calculation of cost and time: first determine the cost and time for all base events, then propagate the cost and time through the base events to the upper events up to the top event. An example is given below: Basic events 'floodplain' and 'IP address masquerading' cost respectively €400 and €400, and required time is 2 weeks and 3 weeks. Then the cost of and('floodplain', 'IP address masquerading') is the the sum(€800) and the time is the sum(5 weeks). Next determine the cost and time of all events in the upper layer('network traffic attack', 'protocol attack', and 'resource exhaustion'), the cost and time of this layer are (€800, 5 weeks), (€500, 4 weeks), and (€1000, 7 weeks), respectively. Then the cost of or ('network traffic attack', 'protocol attack', and 'resource exhaustion') is the min (€500), the time is the 4 weeks. Since the value is propagated from bottom to top, there are two events ('constructing malicious input', 'DoS attack') in the next upper level, and an or gate is used. The set of costs and events is (€500, 4 weeks), (€300, 2 weeks), then the cost and time of or gate is the min (€300, 2 weeks). Finally, the top event has three children nodes , their set of cost and time is (€100, 1 weeks), (€100, 1 weeks), (€300, 2 weeks), then the cost and time of or is the min (€100, 1 weeks).

4.3 Fault tree

Although this research has constructed an attack tree model of chatGPT for understanding the potential attack paths and vulnerabilities in the system, there are also system failure and malfunctioning issues in the chatGPT system that need to be analysed, and these possible points of risk may originate from the user's actions and the flaws in the system itself. For this reason this research will also construct a fault tree model for chatGPT system.

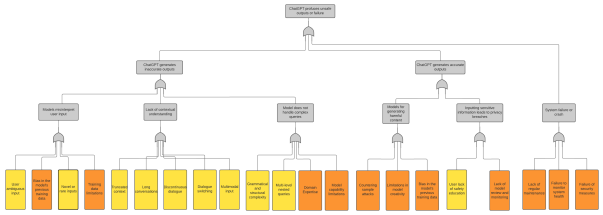

Figure 4: Fault tree yellow basic events represent user-related, red basic events represent system-related

4.3.1 Constructing fault tree. The fault tree model helps to understand the failure mechanism of the system by graphically representing the possible failures of the system and the root causes of the failures. Firstly, the system in this study is chatGPT, and secondly the failure objective of the system is that the chatGPT system does not provide the expected service. Therefore the top event of the failure tree is determined as chatGPT provides unsafe output or the system denies service. Then, the base events are determined based on the top events: chatGPT generates inaccurate outputs, chatGPT generates accurate outputs, and chatGPT system failure or crash.

(1) ChatGPT generates inaccurate outputs
- ChatGPT misinterpret user inputs
  - Ambiguous input
    In everyday use, users may give inaccurate input prompts. Post-processing is occasionally necessary due to imprecise prompts, deviations from the expected behaviour by chatGPT, and its lack of stability [14]. Although post-processing can improve the quality of the output, inappropriate output leaves occur from time to time.
  - Bias in the model's previous training data
    Bias in language modelling refers to the presence of apparent inaccuracies or stereotypes in the output produced by the system based on input cues, which largely depends on the database on which it is trained and reflects the various biases present in the database [31].

Table 1: Analysis of attack tree

Attack Type	Techniques	Resource	Time	Cost	Level
Overconfidence and misinformation	Basic knowledge of NLP and Prompt Engineering techniques	One device	One week	€100	low
Countering Sample Attacks	Expertise in deep learning and adversarial sample attacks	One device	two weeks	€400	medium
Long-term dependency problems	Basic knowledge of NLP and Prompt Engineering	One device	one week	€100	low
Discriminatory output	Technical level of deep learning,natural language processing and model understanding	One device	two weeks	€200	low
Inappropriate content generation	Technical knowledge of deep learning and natural language processing	One device	one week	€100	low
ChatGPT leakage of sensitive information	Technical level of deep learning, natural language processing and model understanding	One device	three weeks	€400	medium
Constructing malicious input	Requires some knowledge of how LLMs work and the ability to craft effective prompts, very well programing skills and Hacking knowledge and tools	One device	three weeks	€200	medium
Network traffic flooding	• Basic Networking Knowledge • Network protocols such as HTTP, TCP, UDP • Attack tools such as DDoS tools • Distributed Systems • Forgery Techniques • Network programming such as socket programming • Basic security	• One device • DDoS tool • programme tool such as python	five weeks	€800	high
Protocol attacks	• Protocol Basics • Protocol Specifications • Vulnerability Analysis • Network capture and analysis such as Wireshark • Programming skills such as python	• One device • network analysis tool such as wireshake • Programme software such as python	four weeks	€500	high
Resource exhaustion	• Understanding the architecture, applications, operating system, and network configuration of the target system Learn about • Known vulnerabilities or security weaknesses that may exist in the target system • Demonstrate penetration testing skills • Scripting language to generate customized attack payloads • Network protocols running on the target system • Encryption and decryption skills • Knowledge of security protocols • Testing tool scripting language such as python	• one or more devices • testing tool • scripting language such as python	seven weeks	€1000	high

These biases may be presented in unpredictable ways to the users in unpredictable ways, resulting in a poor user experience and potentially poor social impact.
  - Novel or rare inputs
    The output of the system depends on the training data, and if the trained data does not contain the prompted input given by the user, the system may not be able to process the input correctly, resulting in the generation of incorrect or inaccurate output.
  - Training data limitations
    Although the amount of available training data is already very large, it is still not enough. This can be seen in the fact that chatGPT is still constantly using users' experience data to improve the system. The limitation of the training data may lead to the generation of outputs that contain incorrect information and mislead users.
- Lack of contextual understanding
  - Truncated context
    If the text entered by the user contains a large amount of information and the model is constrained by the length of the input when processing it, then it may happen that the context is truncated, leading to a lack of full understanding.
  - Long conversations
    Over the course of a long dialogue, the model may gradually forget previous dialogue history, leading to a lack of contextual understanding. This leads to absurd or irrelevant results in the generated output.
  - Discontinuous dialogue
    If the dialogue provided by the user is not a continuous context, but a series of unrelated sentences, the model may not be able to effectively organise this information to capture the full context. Although chatGPT can grasp contextual information, it might endeavor to engage in intricate or multi-level, multi-turn conversations, resulting in unpredictable outcomes that require coherence [32].
  - Dialogue switching
    When a dialogue switches from one topic or context to another, the model may need to adapt to the new context, but if the model fails to understand the context correctly during adaptation, it may result in a lack of full comprehension.
  - Multimodal input
    If the user input contains multiple information types (e.g., text, image, speech) and the model can only process one of those types, it may lack a comprehensive understanding of the overall context.
- ChatGPT does not handle complex queries
  - Grammatical and structural complexity
    Complex syntactic structures or long sentences may make the model difficult to understand and interpret. This may lead to incorrect interpretation or partial understanding of the query.
  - Multi-level nested queries
    If the query contains multiple layers of nested information or covers multiple topics, the model may be confused in its processing.
  - Domain Expertise
    If the query relates to expertise in a particular domain and the model lacks relevant information in that domain, this may result in the model being unable to handle complex specialised queries. The model may generate format-accurate information, such as non-existent references. ChatGPT may misrepresent its own knowledge, choosing to provide fabricated information in what appears to be a highly confident manner, rather than answering "I don't know" [20].
  - Model capability limitations
    Models may not perform well with large queries that exceed their capacity or training range.

(2) ChatGPT generates accurate outputs
- Models for generating harmful content
- Inputting sensitive information leads to privacy breaches

(3) ChatGPT system failure or crash
- Lack of regular maintenance
  Perform regular system maintenance, including software and hardware updates. Ensure that all components of the system are up to date and that potential vulnerabilities have been fixed.
- Failure to monitor system health
  Use monitoring tools and alarm systems to monitor system performance in real time. When anomalies or potential malfunctions are detected, alerts are issued in time for quick response
- Failure of security measures
  Stringent security measures are implemented to guard against potential attacks and malicious behaviour to ensure that the system is not exposed to security threats.

4.3.2 Analysis of fault tree . Fault tree analysis is a methodology used to assess the safety and reliability of a system by graphically representing the possible faults and root causes of failure. In this research, the fault tree will be analysed qualitatively in order to assess the reliability of the system and to identify possible points of improvement. Looking at all the basic events of the fault tree diagram, it is possible to classify all the fault events into two categories: user-caused events and events caused by the system itself.

User-caused failure events: User-caused events are fault events that are caused by unreasonable user actions. For example, the prompts for user inputs are very vague, resulting in the system not being able to process the inputs correctly, so that the system output is inaccurate. According to the fault map, about half of the fault events are caused by user operations. And the interaction between users and the system is mainly the input operation. Human-operated faults can be significantly reduced by improving users' awareness of network security and Prompt Engineering techniques.

System-induced failure events: Events of inaccuracy, unsafe output or system crash due to system limitations. The limitations of the system lie primarily in the training data. Firstly, if there is a bias in the training data, it will be reflected in the output of the

system. Secondly, the training data, although already large, still cannot contain everything. Novel inputs or specialised inputs are most likely to give specious answers. Finally, the chatGPT system, like many other systems, needs to be networked, which leads to the system being exposed to DoS attacks as well.

4.4 Integrated tree

4.4.1 *Constructing attack fault tree.*

- Analyse the events that overlap between the fault tree and the attack tree and find the overlapping part of the events from the top event to the base event.
- Based on the attack tree, add different base events in the fault tree to the attack tree. The principle of adding: from the top event to the base event constitutes the attack path, the base event in the fault tree is compared layer by layer according to the attack path, if it is exactly the same as the attack path, then the attack tree will not be changed, if it is different from the attack path, then the base event of the fault tree will be added from a different place, the fault attack tree will be derived as in Figure 5.

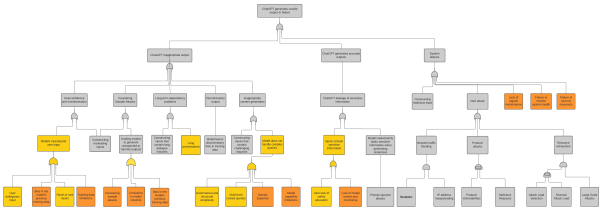

Figure 5: Integrated Tree Yellow basic events represent user-related, red basic events represent system-related

4.4.2 *Analysis of attack-fault tree.*

- There are three types of nodes or base events in the integration tree, namely security nodes, safety nodes, and safety-security nodes. Security nodes are events unique to the attack tree, safety nodes are events unique to the fault tree, and safety-security nodes are events unique to the attack tree, safety nodes are events unique to the fault tree, and safety-security nodes are events common to both the attack and fault trees.
- Cybersafety has a negative impact on cybersecurity. ChatGPT relies heavily on user input and system output to interact. Most safety events can lead to the effects of an attack event, which may result in harmful or inaccurate output from the system, sensitive data leakage, or even system failure. For example, long conversations can be used by users for normal operations or can be exploited by attackers for malicious purposes.
- Security has an indirect positive effect on safety. As shown in the figure 5 some of the safety nodes have consistent or one level lower attack paths than the security nodes, which indicates that either human-caused failures or system-caused failures can be maliciously exploited by an attacker. However, it also shows that increasing the security capability of the system can reduce the frequency of attacks while also reducing the faults. For example, adversarial sample attacks refer to the act of tricking a machine learning model into producing errors in processing these modified inputs by making small but deliberate modifications to the input data. This type of attack aims to exploit the vulnerability of the model by modifying the input data in such a way that it produces misleading results in terms of prediction or classification. Adversarial sample attacks can be a potential threat to security and robustness, as the model may not be able to correctly recognise or classify the modified data when it is processed. If chatGPT takes adversarial training and uses integrated learning both prevent adversarial sample attacks and improve the robustness of the system. The improved robustness of the system also improves the cybersafety of the system.

5 RESULT

1. Identifying where is cybersafety and where is cybersecurity in the chatGPT?

According to fault tree and attack tree to separate cybersecurity and cybersafety

2. How does cybersafety and cybersecurity interact with each other in the chatGPT?

See as Integrated tree and the analysis of Integrated tree.

3. How does AI effect the cybersecurity and cybersafety for the user based on the chatGPT?

AI, including models like chatGPT, can have both positive and negative implications for cybersecurity and cybersafety. Here are some ways in which AI may impact these areas:

- Positive impacts in cybersecurity
  - Threat Detection: AI can be employed to enhance threat detection capabilities. Machine learning models can analyze patterns of normal behavior and identify anomalies that may indicate a security threat or cyberattack [16].
  - Phishing Detection: AI models can be trained to recognize patterns associated with phishing emails and malicious links, aiding in the detection and prevention of phishing attacks [2].
- Negative impacts in cybersecurity
  - Emerging AI-related attacks become more simple and easy, have a low threshold for normal people. For example, phishing attacks and social engineering attacks [26].
  - Users become more passive in the networked environment, and many seemingly normal operations can be hazardous. The integration of AI could enhance the capabilities of existing criminal activities, and there is the potential for the emergence of novel types of crimes that have not been previously identified [10].
  - Adversarial Attacks: Sophisticated attackers may attempt to manipulate AI models, including chatGPT, by feeding them malicious inputs to generate unintended or harmful outputs [33]. This can be a challenge in maintaining the integrity of the system.
  - Automated Cyberattacks: AI can be used by cybercriminals to automate and optimize cyberattacks, making them

more scalable and efficient. This includes automated reconnaissance, vulnerability identification, and exploitation. Adversaries are continually evolving and refining their tactics, placing a significant focus on integrating AI-driven methodologies into their attack strategies. This category of attacks, known as AI-based cyber attacks, leverages artificial intelligence in tandem with traditional attack techniques to amplify the potential for harm and increase the overall impact [12].
  - Social Engineering: ChatGPT and similar models could be used in social engineering attacks to generate convincing and personalized phishing messages, increasing the likelihood of successful attacks. For example, voice spoofing/cloning, deep forgery and automated AI-based socially engineered bots are becoming easier and more difficult to watch out for with the addition of AI [19].
  - Privacy Concerns: The use of AI in cybersecurity often involves the analysis of large amounts of data. Privacy concerns may arise if these analyses involve sensitive or personally identifiable information. ChatGPT engages in conversations with users, and this interaction may unintentionally include the sharing of personal details like names, addresses, contact information, or potentially sensitive records such as medical information. Although the aim is to offer a tailored and interactive experience, there exists a potential risk of inadvertent disclosure or inappropriate storage of such sensitive information during the course of the conversation [8].
- Positive impacts in cybersecurity
  - Incident Response: AI-powered tools can facilitate rapid incident response by automating the analysis of security events, helping security teams to identify and mitigate threats more efficiently [22].
- negative impacts in cybersafety
  - users become more passive in the networked environment, and many seemingly normal operations can be hazardous.
  - Bias in Security Systems: If not carefully designed and monitored, AI models may inherit biases present in training data, leading to discriminatory or unfair outcomes in security-related decisions. From the available literature, it is clear that the chatGPT has been infused with numerous biases since its initial launch [28].

4. In which way is AI useful for the attacker?

- AI, as an increasingly sophisticated and large new field, can provide new ground for cybercrime. This is an area where the laws are not yet robust or adequate and some of the vague cybercrimes are not yet punishable. The way artificial intelligence interacts with cybersecurity is accelerating. In addition to existing challenges, the way AI interacts with cybersecurity is accelerating, creating new security challenges [1].
- AI lowers the skill threshold for cybercriminals, who can use chatGPT to automate parts of their code writing, reducing their reliance on programming and scripting knowledge [26].
- AI such as chatGPT can also help cybercriminals generate fraudulent information to trick people into believing false statements, fictional stories or untrue statements. This helps cybercriminals to commit online fraud and false propaganda. Deepfake technology leverages neural networks and utilizes chatGPT DeepNLP to generate diverse simulated and counterfeit images, making it challenging for forensic analysis to detect the manipulated visuals. This form of technology has been widely employed across different social media platforms, serving as a source of entertainment, with notable examples including instagram and snapchat [4].
- Malicious users may also access sensitive information through chatGPT, which violates the privacy of others and exacerbates cyber fraud [27].

6 CONCLUSIONS

In conclusion, the research finds that cybersafety negatively impacts cybersecurity in chatGPT due to its reliance on user input and system output. However, enhancing security capabilities can reduce faults and attack frequency, illustrating that security indirectly benefits safety. The research underscores the complex interplay between cybersecurity, cybersafety, and AI, highlighting the need for robust and integrated strategies to manage these interactions effectively.

REFERENCES

[1] Meraj Farheen Ansari, Bibhu Dash, Pawankumar Sharma, and Nikhitha Yathiraju. 2022. The impact and limitations of artificial intelligence in cybersecurity: a literature review. *International Journal of Advanced Research in Computer and Communication Engineering*.

[2] Abdul Basit, Maham Zafar, Xuan Liu, Abdul Rehman Javed, Zunera Jalil, and Kashif Kifayat. 2021. A comprehensive survey of AI-enabled phishing attacks detection techniques. *Telecommunication Systems*, 76, 1, (Jan. 1, 2021), 139–154. DOI: 10.1007/s11235-020-00733-2.

[3] Dan Craigen, Nadia Diakun-Thibault, and Randy Purse. 2014. Defining cybersecurity. *Technology Innovation Management Review*, 4, 10, 13–21. Place: Ottawa Publisher: Talent First Network.

[4] Bibhu Dash and Pawankumar Sharma. 2023. Are chatgpt and deepfake algorithms endangering the cybersecurity industry? a review. *International Journal of Engineering and Applied Sciences*, 10, 1.

[5] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. 2023. Not what you've signed up for: compromising real-world LLM-integrated applications with indirect prompt injection. (May 5, 2023). Retrieved Nov. 24, 2023 from http://arxiv.org/abs/2302.12173 arXiv: 2302.12173[cs].

[6] Richard J. Harknett and James A. Stever. 2011. The new policy world of cybersecurity. *Public Administration Review*, 71, 3, 455–460. _eprint: https://onlinelibrary.wiley.com/doi/pdf 6210.2011.02366.x. DOI: 10.1111/j.1540-6210.2011.02366.x.

[7] Matthias Hofstetter, Reinhard Riedl, Thomas Gees, Adamantios Koumpis, and Thomas Schaberreiter. 2020. Applications of AI in cybersecurity. In *2020 Second International Conference on Transdisciplinary AI (TransAI)*. 2020 Second International Conference on Transdisciplinary AI (TransAI). (Sept. 2020), 138–141. DOI: 10.1109/TransAI49837.2020.00031.

[8] Ken Huang, Fan Zhang, Yale Li, Sean Wright, Vasan Kidambi, and Vishwas Manral. 2023. Security and privacy concerns in ChatGPT. In *Beyond AI: ChatGPT, Web3, and the Business Landscape of Tomorrow*. Future of Business and Finance. Ken Huang, Yang Wang, Feng Zhu, Xi Chen, and Chunxiao Xing, (Eds.) Springer Nature Switzerland, Cham, 297–328. ISBN: 978-3-031-45282-6. DOI: 10.1007/978-3-031-45282-6_11.

[9] Sajed Jalil, Suzzana Rafi, Thomas D. LaToza, Kevin Moran, and Wing Lam. 2023. ChatGPT and software testing education: promises & perils. In *2023 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*. (Apr. 2023), 4130–4137. arXiv: 2302.03287[cs]. DOI: 10.110 9/ICSTW58534.2023.00078.

[10] Doowon Jeong. 2020. Artificial intelligence security threat, crime, and forensics: taxonomy and open issues. *IEEE Access*, 8, 184560–184574. Conference Name: IEEE Access. DOI: 10.1109/ACCESS.2020.3029280.

[11] Chris Johnson. 2012. CyberSafety: CyberSecurity and safety-critical software engineering. In *Achieving Systems Safety*. Chris Dale and Tom Anderson, (Eds.)

Springer, London, 85–95. ISBN: 978-1-4471-2494-8. DOI: 10.1007/978-1-4471-2494-8_8.

[12] Nektaria Kaloudi and Jingyue Li. 2020. The AI-based cyber threat landscape: a survey. *ACM*, 53, 1, (Feb. 6, 2020), 20:1–20:34. DOI: 10.1145/3372823.

[13] Shaharyar Khan and Stuart Madnick. 2022. Cybersafety: a system-theoretic approach to identify cyber-vulnerabilities & mitigation requirements in industrial control systems. *IEEE Transactions on Dependable and Secure Computing*, 19, 5, (Sept. 2022), 3312–3328. Conference Name: IEEE Transactions on Dependable and Secure Computing. DOI: 10.1109/TDSC.2021.3093214.

[14] Jan Kocoń et al. 2023. ChatGPT: jack of all trades, master of none. *Information Fusion*, 99, (Nov. 1, 2023), 101861. DOI: 10.1016/j.inffus.2023.101861.

[15] Barbara Kordy, Ludovic Piètre-Cambacédès, and Patrick Schweitzer. 2014. DAG-based attack and defense modeling: don't miss the forest for the attack trees. *Computer Science Review*, 13-14, (Nov. 1, 2014), 1–38. DOI: 10.1016/j.cosrev.2014.07.001.

[16] Jonghoon Lee, Jonghyun Kim, Ikkyun Kim, and Kijun Han. 2019. Cyber threat detection based on artificial neural networks using event profiles. *IEEE Access*, 7, 165607–165626. Conference Name: IEEE Access. DOI: 10.1109/ACCESS.2019.2953095.

[17] Bowen Liu, Boao Xiao, Xutong Jiang, Siyuan Cen, Xin He, and Wanchun Dou. 2023. Adversarial attacks on large language model-based system and mitigating strategies: a case study on ChatGPT. *Security and Communication Networks*, 2023, (June 10, 2023), e8691095. Publisher: Hindawi. DOI: 10.1155/2023/8691095.

[18] Kevin Macnish and Jeroen van der Ham. 2020. Ethics in cybersecurity research and practice. *Technology in Society*, 63, (Nov. 1, 2020), 101382. DOI: 10.1016/j.techsoc.2020.101382.

[19] Sowjanya Manyam. [n. d.] Artificial intelligence's impact on social engineering attacks.

[20] Jesse G. Meyer et al. 2023. ChatGPT and large language models in academia: opportunities and challenges. *BioData Mining*, 16, 1, (July 13, 2023), 20. DOI: 10.1186/s13040-023-00339-9.

[21] Stefano M. Nicoletti, Marijn Peppelman, Christina Kolb, and Mariëlle Stoelinga. 2023. Model-based joint analysis of safety and security:survey and identification of gaps. *Computer Science Review*, 50, (Nov. 1, 2023), 100597. DOI: 10.1016/j.cosrev.2023.100597.

[22] Constantin Nilă, Ioana Apostol, and Victor Patriciu. 2020. Machine learning approach to quick incident response. In *2020 13th International Conference on Communications (COMM)*. 2020 13th International Conference on Communications (COMM). (June 2020), 291–296. DOI: 10.1109/COMM48946.2020.9141989.

[23] Ramjee Prasad and Vandana Rohokale. 2020. Artificial intelligence and machine learning in cyber security. In *Cyber Security: The Lifeline of Information and Communication Technology*. Springer Series in Wireless Technology. Ramjee Prasad and Vandana Rohokale, (Eds.) Springer International Publishing, Cham, 231–247. ISBN: 978-3-030-31703-4. DOI: 10.1007/978-3-030-31703-4_16.

[24] Partha Pratim Ray. 2023. ChatGPT: a comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. *Internet of Things and Cyber-Physical Systems*, 3, (Jan. 1, 2023), 121–154. DOI: 10.1016/j.iotcps.2023.04.003.

[25] John Sammons and Michael Cross. 2016. *The Basics of Cyber Safety: Computer and Mobile Device Safety Made Easy*. Google-Books-ID: vLNZAwAAQBAJ. Elsevier, (Aug. 20, 2016). 255 pp. ISBN: 978-0-12-416639-4.

[26] Glorin Sebastian. 2023. Do ChatGPT and other AI chatbots pose a cybersecurity risk?: an exploratory study. *International Journal of Security and Privacy in Pervasive Computing*, 15, 1, (Mar. 22, 2023), 1–11. DOI: 10.4018/IJSPPC.320225.

[27] Glorin Sebastian. 2023. Privacy and data protection in ChatGPT and other AI chatbots: strategies for securing user information. *SSRN Electronic Journal*. DOI: 10.2139/ssrn.4454761.

[28] Sahib Singh. 2023. Is chatgpt biased? a review.

[29] Mariarosaria Taddeo, Tom McCutcheon, and Luciano Floridi. 2019. Trusting artificial intelligence in cybersecurity is a double-edged sword. *Nature Machine Intelligence*, 1, 12, (Dec. 2019), 557–560. Number: 12 Publisher: Nature Publishing Group. DOI: 10.1038/s42256-019-0109-1.

[30] the Universiti Kebangsaan Malaysia, Malaysia, N. A. A Rahman, I. H. Sairi, N. A. M. Zizi, and F. Khalid. 2020. The importance of cybersecurity education in school. *International Journal of Information and Education Technology*, 10, 5, 378–382. DOI: 10.18178/ijiet.2020.10.5.1393.

[31] Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., 12388–12401. Retrieved Jan. 15, 2024 from https://proceedings.neurips.cc/paper/2020/hash/92650b2e92217715fe312e6fa7b90d82-Abstract.html.

[32] Taylor Webb, Keith J. Holyoak, and Hongjing Lu. 2023. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7, 9, (Sept. 2023), 1526–1541. Number: 9 Publisher: Nature Publishing Group. DOI: 10.1038/s41562-023-01659-w.

[33] Wei Emma Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. 2020. Adversarial attacks on deep-learning models in natural language processing: a survey. *ACM Transactions on Intelligent Systems and Technology*, 11, 3, (Apr. 3, 2020), 24:1–24:41. DOI: 10.1145/3374217.