

# Integrating Cognitive Work Analysis into an ACT-R Model for Cybersecurity Applications

by

Fan He

A thesis

presented to the University of Waterloo

in fulfillment of the

thesis requirement for the degree of

Doctor of Philosophy

in

Systems Design Engineering

Waterloo, Ontario, Canada, 2025

© Fan He 2025

## Examining Committee Membership

The following served on the Examining Committee for this thesis. The decision of the Examining Committee is by majority vote.

**External Examiner:**

*B.L. William Wong*

Professor, Dept. of Computer Science and Software Engineering

Auckland University of Technology, New Zealand

**Supervisor:**

*Catherine M. Burns*

Professor, Dept. of Systems Design Engineering

University of Waterloo

**Internal Member:**

*Shi Cao*

Associate Professor, Dept. of Systems Design Engineering

University of Waterloo

**Internal Member:**

*Sebastian Fischmeister*

Professor, Dept. of Electrical and Computer Engineering

University of Waterloo

**Internal-External Member:**

*Adam Molnar*

Assistant Professor, Dept. of Sociology & Legal Studies

University of Waterloo

## **Author's Declaration**

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

## Statement of Contributions

Fan He was the sole author for Chapters 1, 2 and 7, which were written under the supervision of Dr. Catherine Burns and were not written for publication.

This thesis consists of two manuscripts written for publication and one conference abstract that has been presented. Exceptions to sole authorship of material are as follows:

### **Research presented in Chapter 3:**

This research was conducted at the University of Waterloo by Fan He under the supervision of Dr. Catherine Burns and Dr. Sebastian Fischmeister. Fan He and Karim Elhammady were responsible for data collection and analysis. Fan He wrote the draft manuscripts, which all co-authors contributed intellectual input on.

He, F., Elhammady, K., Fischmeister, S., & Burns, C. M. (2024, May). Preliminary Cognitive Modeling: Comparing Distraction-Based Cyber-Attacks and Alcohol-Related Impairments on Human Drivers. In 2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS) (pp. 1-6). IEEE.

### **Research presented in Chapter 4:**

This research was conducted at the University of Waterloo by Fan He under the supervision of Dr. Catherine Burns. Fan He drafted the abstract, and each co-author provided intellectual input on its revisions.

He, F., Cao, S., J., Fischmeister, S., & Burns, C.M., (2024). Integrating Cognitive Work Analysis (CWA) with Adaptive Control of Thought (ACT-R). *ASPIRE - 68<sup>th</sup> HFES International Annual Meeting*. Phoenix, Arizona, U.S.

### **Research presented in Chapters 5:**

This research was conducted at the University of Waterloo by Fan He under the supervision of Dr. Catherine Burns and Dr. Sebastian Fischmeister. Fan He and Juliet Kern contributed to the experimental design, prototyping, data collection, and analysis. Fan He wrote the draft manuscripts on which all co-authors contributed intellectual input.

He, F., Kern, J., Fischmeister, S., & Burns, C. M. (2024, September). Awareness and Assistance: General Drivers' Cyber Threat Identification and the Role of an In-Vehicle Console Display. In 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC) (pp. 3607-3612). IEEE.

### **Research presented in Chapters 6:**

This research was conducted at the University of Waterloo by Fan He under the supervision of Dr. Catherine Burns. Fan He contributed to the experimental design and prototyping. Fan He completed the interviews, data collection, and analysis with assistance from Hong Zhang and Rachel Baek.

As lead author of these four chapters, I was responsible for contributing to conceptualizing study design, carrying out data collection and analysis, and drafting and submitting manuscripts. My coauthors provided guidance during each step of the research and provided feedback on draft manuscripts.

## Abstract

Cybersecurity is a trending concern with the rapid development of many systems. While humans are often considered vulnerable targets, research on human factors remains limited compared to the extensive technical focus on defense and mitigation strategies. Human-focused cognitive research in this domain faces two primary challenges: the evolving and complex nature of the cybersecurity landscape, and the domain-specific characteristics of the systems under attack. These challenges point to the need for modeling human performance in identifying vulnerabilities, with both precise dynamic measurement and domain-specific fidelity.

Accordingly, we proposed a solution by integrating CWA into ACT-R models. A detailed elaboration on the CWA and ACT-R's structural compatibility across dimensions, their fundamental strengths as complements, and the functional competencies with integration was presented. This conceptual exploration demonstrated the feasibility of integrating the CWA and ACT-R, leading to improvements in model construction efficiency and domain-specific validity.

We explored CWA and ACT-R for modeling humans in vehicle cybersecurity. While we were able to demonstrate a model, a follow-up study with human participants showed that drivers may not actively identify vulnerabilities and mitigate cyber threats. We then practically implemented and applied the integrated model, from model construction preparation to detailed rule development, guided by CWA's Work Domain Analysis, Control Task Analysis, and Strategies Analysis, to simulate the SOC analysts' cybersecurity alert triage performance. The model construction process demonstrated better efficiency with a systematic approach, and the resulting model showed improvement trend in quantitative accuracy, domain-specific validity, and the interpretability of human adaptability and flexibility. However, the model is limited in capturing human exploratory behavior, prompting a brief test of using Generative AI (GAI) models to address this gap.

This thesis is the first exploration and implementation of integrating CWA-guided domain-specific analysis with ACT-R's computational capabilities to develop an integrated cognitive model for humans in complex work domains. The effort advances the development of cognitive modeling by providing theoretical grounding and practical insights for applying and extending cognitive models. Finally, we discuss whether GAI models might enhance cognitive modeling, as GAI capabilities become more available.

## Acknowledgements

I want to begin by thanking my supervisor, Dr. Catherine Burns, for her continuous support throughout these years. She has fostered a research environment full of trust, creativity, and encouragement that allowed me to enjoy and grow throughout this journey.

I would also like to thank my committee members: Dr. Sebastian Fischmeister, Dr. Shi Cao, Dr. Adam Molnar, and Dr. B.L. William Wong, for their guidance and valuable feedback, which greatly contributed to the improvement of this work.

I want to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC) and the Maks Wulkan Graduate Scholarship for providing financial support during my PhD.

My sincere thanks to Dr. Yeti Li for his generous help, and great vision in jumpstarting my research journey. Thanks to CNL human factors scientists for their mentorship during my internship, and to all current and former AIDL lab members for their company along the way.

Finally, I would like to thank my parents for their unconditional love and trust. Thanks to Hugo for his unique sense of humor. This work would not have been possible without them.

# Table of Contents

Author’s Declaration.....	iii
Statement of Contributions .....	iv
Abstract .....	vi
Acknowledgements.....	vii
List of Figures .....	xiv
List of Tables.....	xvi
List of Abbreviations.....	xvii
Chapter 1 Introduction .....	1
1.1 Background .....	2
1.1.1 Defining Cybersecurity and Related Concepts .....	2
1.2 Literature Review on Cognitive Modeling in Cybersecurity .....	5
1.2.1 Task Characteristics in Cybersecurity for Cognitive Modeling.....	7
Chapter 2 Cognitive Models: Challenges in Modeling Human Performance in Cybersecurity .....	10
2.1 Cognitive Architectures and Models.....	10
2.1.1 Review of Cognitive Modeling: Applications with Different Emphases .....	12
2.2 Adaptive Control of Thought-Rational (ACT-R) .....	14
2.2.1 ACT-R: Strategies by Symbolic Construction and Sub-symbolic Selection .....	16
2.3 Research Questions .....	17
2.4 Thesis Structure.....	20
2.5 Contributions.....	23
Chapter 3 An Example of the ACT-R Model Application in Cybersecurity .....	24
3.1 ACT-R Models Strengths in CAV Cybersecurity.....	24
3.2 Distraction-Based Cyberattacks Targeting In-Vehicle Human Operations .....	25

3.2.1 Assumptions .....	26
3.2.2 Model Construction.....	27
3.2.3 Model’s Simulation Results and Evaluations .....	29
3.3 Discussions of Applying ACT-R Model in Cybersecurity .....	30
3.3.1 Insufficient Systematic Analysis to Support Rule Construction and Mapping.....	31
3.3.2 Generalized Task Environment.....	33
3.3.3 Defining the Effective Task Environment of the Modeling’s Application.....	35
3.4 Future Work .....	35
3.5 Summary of Chapter 3 .....	36
Chapter 4 Towards Integrating CWA with the ACT-R Model.....	38
4.1 Cognitive Architecture, Model, and Analysis Framework .....	38
4.2 Introduction to Cognitive Work Analysis (CWA).....	39
4.3 Towards an Enhanced Model through the Integration of CWA .....	40
4.3.1 Conceptual Overview of the Integration .....	41
4.3.2 Abstraction Hierarchy (AH) and Declarative Knowledge .....	41
4.3.3 From Decision Ladder (DL) to Production Rules.....	43
4.3.4 Strategy Analysis and Production Rules Selection .....	45
4.3.5 Overview of the Integration .....	46
4.4 Contribution .....	46
4.5 Summary .....	47
Chapter 5 Cognitive Model Application: Assessing In-Vehicle Human Responses to CAV Cyberattacks.....	49
5.1 Study 1: Driver Responses to Silent and Explicit Cyberattacks with In-Vehicle Display Interfaces .....	49

5.1.1 Introduction .....	49
5.1.2 Background .....	50
5.1.3 Hypotheses .....	52
5.1.4 Experiment Design.....	52
5.1.5 Methods.....	54
5.1.6 Results Analysis .....	56
5.2 Discussions.....	60
5.2.1 Validation of Study 1 Hypotheses.....	60
5.2.2 Challenges in Modeling In-Vehicle Human Responses.....	62
5.2.3 Modeling Task Shift Toward Security Operations Centers (SOCs) .....	63
5.3 Limitations .....	64
5.4 Summary .....	65
Chapter 6 Study 2: Predicting SOC Analysts' Alert Triage Task Performance with a Cognitive Model Integrating CWA and ACT-R.....	67
6.1 Overview .....	67
6.2 Background .....	68
6.2.1 Modeling Focus Shift from In-Vehicle Operations to SOC.....	68
6.2.2 Internship Experience and Observations.....	68
6.3 Introduction to Cybersecurity Operation Center Domain .....	69
6.3.1 Literature Review: Cybersecurity Operation Center Research .....	69
6.4 Applying CWA in Modelling SOC Alert Triage .....	74
6.4.1 WDA and Abstraction Hierarchy.....	74
6.4.2 Methods for Work Domain Analysis (WDA).....	75
6.4.3 Settings and AH Modelling.....	76

6.4.4	Conducting WDA.....	77
6.4.5	Translating to Interface Design Elements .....	82
6.4.6	Alert Triage Tool Interface Development.....	83
6.4.7	Experiment Design.....	86
6.4.8	Defined vs. Ambiguous Pattern Alerts.....	88
6.4.9	Methods for Control Task Analysis (ConTA) .....	88
6.4.10	Decision Ladder (DL) of Tier-1 Analysts Alert Triage Task .....	89
6.4.11	Methods for Strategy Analysis.....	90
6.4.12	Strategy Analysis to Information Flow Map.....	91
6.5	ACT-R Models Construction .....	96
6.5.1	The Basic Model .....	96
6.5.2	Strategy Analysis into Production Rules.....	97
6.5.3	Control Task Analysis to Production Map.....	98
6.5.4	CWA to Declarative Modules.....	101
6.5.5	Overview of Differences: CWA-Informed Model vs. Basic Model.....	101
6.5.6	Global Parameters .....	103
6.5.7	Rewarding Mechanism and Utility Manipulation.....	103
6.5.8	Exclusion of Adaptive Strategy Modeling and Ambiguous Pattern Alerts .....	105
6.6	Results .....	105
6.6.1	Participants' Behavioral Data Analysis.....	105
6.6.2	Human Data vs. Models Simulation .....	107
6.7	Discussions.....	109
6.7.1	Model Computational Performance: Human vs. Simulation Results .....	109

6.7.2 Enhanced Model Construction and Application Through the Integration of CWA and ACT-R.....	111
6.7.3 Interface Design and Cognitive Modelling .....	112
6.7.4 Modelling Scope: Modeling Knowledge-Based Decision-Making .....	114
6.7.5 Aspects Not Captured by the Model .....	114
6.8 Conclusions .....	116
6.9 Limitations and Future Works .....	116
6.10 Summary .....	117
Chapter 7 Summary and Future Works.....	119
7.1 Summary of Key Outcomes .....	119
7.2 Enhancements by CWA-Informed ACT-R Models in Cybersecurity .....	122
7.2.1 Cognitive Modeling Informed by Task Knowledge .....	122
7.2.2 Cognitive Modeling with Enhanced Task Environment Fidelity .....	123
7.2.3 Cognitive Modeling with Further Development for Cybersecurity Applications .....	124
7.3 Development of Cognitive Models .....	125
7.3.1 Integration of CWA and ACT-R.....	125
7.3.2 Best Use Case of the Model .....	127
7.3.3 Limitations and Future Development.....	130
7.4 Guidance on the Application of the Integrated Model.....	133
7.5 Contributions.....	136
7.6 Conclusions .....	137
References .....	138
Appendix A .....	174

An Example of Simulating Distraction-Based Cyberattacks Targeting In-Vehicle Human Operations using an ACT-R Based Model.....	174
Appendix B .....	176
Study 2 Information Requirements List Based on the SOC AH Model (Full List) .....	176
Study 2 Decision Ladder Construction Semi-Structured Interview Questions.....	178
Study 2 Alerts List .....	180
Study 2 Supplementary Materials: Runbook (FAQ).....	181
Study 2 Participant List.....	185
Study 2 Production Rules Descriptions.....	186
Study 2 Mapping of Production Rules by Strategy Analysis and ConTA (Individual View of Different Strategies).....	191
Appendix C .....	193
A Follow-Up Study Extending the Integrated Framework: An exploration of GAI Models in Cognitive Modeling .....	193
Follow-up Study Objectives.....	194
Experiment Settings .....	194
Results .....	195
Discussions.....	197
Limitations .....	199
Summary of the Follow-up Study Findings .....	199
Classic Cognitive Models vs. GAI Models.....	200

## List of Figures

Figure 1. The SAE Levels (SAE International, 2021a).....	4
Figure 2. Structure of a Model based on a Cognitive Architecture (Byrne, 2012).....	11
Figure 3. A part of the graphical representation of the practical applications of the cognitive architectures and corresponding competency areas defined in (Adams et al., 2012)..	12
Figure 4. The Basic ACT-R Architecture. ....	15
Figure 5. Overview of Thesis Structure. ....	22
Figure 6. The Basic Task Models from (Deng, Cao, et al., 2019a) and the Reorganization of these Task Models for the Modeling Test. ....	28
Figure 7. The Comparisons of Simulated Reaction Time vs. Reference Values. ....	30
Figure 8. Structure of a Model based on a Cognitive Architecture (Byrne, 2012) (focused on Task Knowledge). ....	32
Figure 9. Structure of a Model based on a Cognitive Architecture (Byrne, 2012) (focused on Task Environment).....	34
Figure 10. Cognitive Architecture, Model, and Framework and How They Function within Human Performance Modeling.....	39
Figure 11. Five Phases of CWA (Rauffet et al., 2015). ....	41
Figure 12. The Abstraction Hierarchy of a vehicle system. ....	42
Figure 13. Declarative Memory Chunk Example (from an Abstraction Hierarchy of a vehicle system). ....	43
Figure 14. A Decision Ladder (DL). ....	44
Figure 15. Comparison of ConTA and Strategies Analysis (Vicente, 1999). ....	45
Figure 16. CWA’s Key Dimensions within ACT-R Architecture. ....	47
Figure 17. The Cognitive Model by the Integration of CWA and ACT-R. ....	48
Figure 18. Example preview of the Tesla Model 3 interface prototype employed in this study. ....	53
Figure 19. Screenshots of Experiment Videos of Cyber Attacks.....	55
Figure 20. Anomaly Detection and Attack Identification Accuracy Across Three Scenarios.....	58
Figure 21. The vehicle system shares attack information with the VSOC (Hamad et al., 2024) .....	64
Figure 22. Study 2 Overview: Integration Dimensions and Experimental Design.....	67
Figure 23. Typical SOC analyst tier responsibilities (Kokulu et al., 2019). ....	72
Figure 24. Typical SOC data and tools (Knerler et al., 2023). ....	78

Figure 25. Basic SOC workflow (Knerler et al., 2023).....	79
Figure 26. The AH work domain model for a SOC (on Tier-1 analyst work scope).....	80
Figure 27. Systematic Approach to Graphic Form Design (C. M. Burns & Hajdukiewicz, 2017). ....	83
Figure 28. Splunk SOAR: Alert List view .....	84
Figure 29. Experiment Alert Triage Tool: Alert-List View .....	84
Figure 30. Experiment Alert Triage Tool: Pop-up Detail.....	85
Figure 31. Decision Ladder for SOC Alert Triage Task.....	92
Figure 32. A screenshot of Alert List view with main indicators .....	94
Figure 33. Pattern-based & Streamlined Strategies (Up); Adaptive Strategy (Bottom). ....	96
Figure 34. A Brief View of the Basic Model Rule Map. ....	97
Figure 35. Example of Pattern-Based Strategy Analysis and Corresponding Production Rules.....	97
Figure 36. Example of Translating Strategy Analysis into a Production Rule Construction.....	98
Figure 37. Example of how Production Rules are connected under the guidance of DL transition. ....	99
Figure 38. Decision Ladder translated to the Production Map .....	100
Figure 39. Structural Comparison of the Two Models and Rule Branches in the CWA-Informed Model Based on Strategy Analysis. ....	102
Figure 40. Differences in production rule connections between the Basic Model and the CWA- informed Model.....	103
Figure 41. Per Alert Processing Time (Defined Pattern vs. Ambiguous Pattern).....	106
Figure 42. Human vs. Simulation Processing Time per Alert. ....	108
Figure 43. Models Agreement with Human Decisions. ....	108
Figure 44. From WDA to ACT-R Setup Preparations. ....	113
Figure 45. Overview of Thesis Structure – Chapter 7.....	119
Figure 46. Between CWA and ACT-R Model Construction. ....	127
Figure 47. Task Environment for the Integrated Model's Applicability. ....	128
Figure 48. Production Map for only Pattern-based Strategy Rules.....	191
Figure 49. Production Map for only Streamlined Strategy Rules. ....	192
Figure 50. Features Selection (key deciding features (Left) vs. supporting evidence features (Right)) between Human Participants, and ChatGPT-4.....	196

## List of Tables

Table 1. The Difference between CVs and AVs (Federal Highway Administration, 2018). .....	4
Table 2. Comparison of Leading Cognitive Architecture and Related Competency Areas. ....	13
Table 3. The "Explicitness" of the Attacks in Scenarios.....	54
Table 4. Console Display Interface Improvement Suggestions. ....	59
Table 5. Chosen Initial Action(s) If Encountering the Anomalies.....	59
Table 6. The Mapping between a subset of WDA-derived information for Declarative Chunk Preparation (see the Full list in Appendix B Table 14).....	85
Table 7. The declarative memory chunk ( <i>alert_position</i> ), used by the model to store spatial and alert- related feature information. ....	101
Table 8. An Overview of the Basic Model vs. the CWA-informed Model. ....	102
Table 9. Specific non-default Utility Assignment.....	104
Table 10. Integration of CWA and ACT-R with Enhancements. ....	116
Table 11. An Integration of CWA and ACT-R with Achieved Enhancement.....	126
Table 12. Guidance on the Application of the Integrated Model.....	135
Table 13. Reference Values for Baseline Reaction Time and Impaired Performance from BAC Study. .....	175
Table 14. Information Requirements Based on the SOC AH Model.....	176
Table 15. Alerts Selected for Study 3. ....	180
Table 16. Recommendations from Runbook. ....	182
Table 17. Study 3 Participants (Group 1) List. ....	185
Table 18. Production Rules Descriptions.....	186
Table 19. The comparison of majority decisions among human, integration of CWA & ACT-R and ChatGPT-4 results. ....	195

## List of Abbreviations

<b>ADAS</b>	Advanced Driver Assistance Systems
<b>HMIs</b>	Human-Machine Interfaces
<b>SOCs</b>	Security Operations Centers
<b>CAV</b>	Connected and Automated Vehicle
<b>CWA</b>	Cognitive Work Analysis
<b>ACT-R</b>	Adaptive Control of Thought-Rational
<b>WDA</b>	Work Domain Analysis
<b>AH</b>	Abstraction Hierarchy
<b>ConTA</b>	Control Task Analysis
<b>EID</b>	Ecological Interface Design
<b>V2X</b>	Vehicle to Everything
<b>V2V</b>	Vehicle to Vehicle
<b>V2I</b>	Vehicle to Infrastructure
<b>SIEM</b>	Security Information and Event Management
<b>SOAR</b>	Security Orchestration, Automation, and Response
<b>DM</b>	Declarative Modules
<b>IOC</b>	Indicators of Compromise
<b>ITS</b>	Intelligent Transportation Systems

# Chapter 1

## Introduction

Many complex socio-technical and safety-critical systems, including those in transportation, healthcare, and the energy domains, have experienced transformative technological advancements and now face growing cybersecurity concerns (Ayodeji et al., 2023; X. Sun et al., 2022; Tariq et al., 2025). These systems now rely more heavily on sensing components, control units, and connectivity mechanisms, introducing expanded targets for cyber threats. Extensive technical research on cyberattacks has covered diverse attack surfaces across domains (Abosata et al., 2021; K. Kim et al., 2021; Sheehan et al., 2019); meanwhile, defense strategies have been actively researched to mitigate impacts and strengthen overall system security (Pham & Xiong, 2020; Tsiknas et al., 2021). In recent years, artificial intelligence techniques with big data analysis have also shown promise in strengthening the systems' ability to detect and manage threats (Illiashenko et al., 2023). However, the consensus is that a perfectly secure system is unattainable (X. Sun et al., 2022). This unattainability stems not only from the ever-evolving nature of the technical landscape but also from the variability of humans' interactions with the system.

Humans are continuously regarded as the most vulnerable and least controllable aspect in cybersecurity (Alsharif et al., 2022; Linkov et al., 2019). Yet, research on the human aspect remains limited compared to technical research in the cyberspace (Maalem Lahcen et al., 2020). Incidents over the past decade have demonstrated the real risks posed by cyber threats across multiple domains (see (Harry & Gallagher, 2025) for a list of reported cyber events from 2014 to 2023). The public has thus become more aware of the potential severity of cyber threats, but is still largely unaware of the detailed vulnerabilities and defensive measures (Alawadhi et al., 2020; Beckers et al., 2022; Zhai et al., 2024). This gap in turn leaves the general population anxious yet unprepared for the emerging cyber threats, exacerbating the risks associated with human vulnerabilities and weakening the effectiveness of the system's cybersecurity safeguards (Mwanje et al., 2024; Oladimeji et al., 2023). Consequently, there is a growing need to proactively support human roles in defending against cyber threats to break this cycle. However, the complexity of the cybersecurity domain and the evolving nature of human roles in these advanced systems together contribute to the challenges of researching human vulnerabilities in the cybersecurity landscape.

Cognitive modeling has proven to be a promising approach for analyzing human behavior and identifying vulnerabilities in the general cybersecurity domain (Maalem Lahcen et al., 2020;

Veksler et al., 2018). Its modelling results could elaborate on the human thinking processes and behaviors, from perception to final decision-making (B. Wu et al., 2022). In this way, cognitive modelling can reveal human errors, limitations, and incapacities that lead to vulnerabilities and further guide the design of human-centric defense solutions targeting the identified weaknesses. Notably, while cybersecurity is itself a dynamic, multifaceted, and ever-evolving domain, it often functions as a supportive layer within more complex systems. Since the characteristics of the work domain also significantly influence human cognitive processes, an effective cognitive modeling needs to incorporate the analysis of the context of specific domain systems (Gersh et al., 2005). Accordingly, applying cognitive modeling in cybersecurity requires tailoring to domain-specific complexities and constraints (Coventry & Branley, 2018; Krause et al., 2021; Linkov et al., 2019; Van Der Kleij et al., 2022).

Therefore, this work aims to leverage current cognitive modeling efforts in cybersecurity with enhanced domain-specificity to identify potential human vulnerabilities and to provide insights for effective mitigation. This work will also elaborate on the challenges for developing and applying a domain-specific cognitive model within dynamic and evolving complex sociotechnical systems. We begin with a modeling focus on CAV systems as an illustrative example. This allows us to examine the challenges of applying cognitive modeling in fast-developing complex domains. The work will then extend the efforts to broader cybersecurity applications for advancing cognitive modeling approaches.

The above provides an introduction of the thesis's motivation and objectives. The following background section introduces key concepts underpinning this thesis, outlines current work and ongoing challenges in cybersecurity cognitive modeling research, and leads to the formulation of our research questions.

## **1.1 Background**

### **1.1.1 Defining Cybersecurity and Related Concepts**

This section first elaborates on definitions of the key concepts in this work.

The National Institute of Standards and Technology (NIST) (Special Publication 800-30, Rev. 1) (Joint Task Force Transformation Initiative, 2012) defines *cybersecurity* as "the process of

protecting information by preventing, detecting, and responding to attacks." The CNSSI (Dukes, 2016) explains cybersecurity-related terms as:

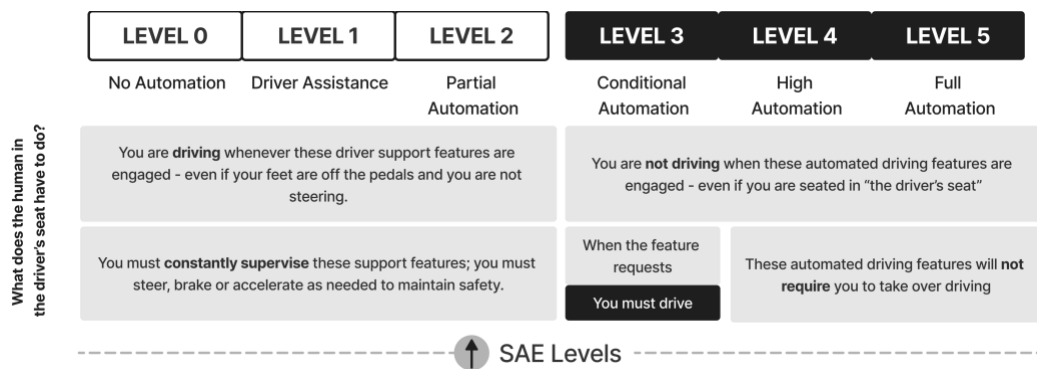
- **Cybersecurity:** The ability to protect or defend the use of cyberspace from cyberattacks.
- **Cyberspace:** A global domain within the information environment, consisting of interdependent networks of information system infrastructures, including the Internet, telecommunications networks, computer systems, and embedded processors and controllers.
- **Cyber Attack:** An attack via cyberspace targeting an enterprise's use of cyberspace to disrupt, disable, destroy, or maliciously control a computing environment/infrastructure, or to destroy data integrity or steal controlled information.

In a broader context, *cybersecurity* on the Internet of Things (IoT) is defined as "the organization and the protection of information technologies with the combination of the following notions: availability, confidentiality, criticality, attack impact, integrity, ownership, sensitive values, legal risk, contextualization, risk assessment and information storage" (Collard et al., 2017). In the transportation domain, the National Highway Traffic Safety Administration (NHTSA, 2021) defines *cybersecurity* as the protection of vehicle components, infrastructure, and communications from harmful attacks, unauthorized access, or any other actions that jeopardize safety functions.

Since we used the CAV system as a potential example to illustrate domain-specific challenges in cognitive modeling for cybersecurity, we also introduce key CAV domain concepts here. The most recognized and comprehensive definitions in the autonomous and automated vehicles domain are from (SAE International, 2018, 2021b).

- ***Automated Vehicles (AVs):*** A given vehicle may be equipped with a driving automation system that is capable of delivering multiple driving automation features that operate at different levels; thus, the level of driving automation exhibited in any given instance is determined by the feature(s) engaged. As such, the recommended usage for describing a vehicle with driving automation capability is "Level [1 or 2] driving automation system-equipped vehicle" or "Level [3, 4, or 5] ADS-equipped vehicle." (See Figure 1 for SAE Automation Levels.)
- ***Automated Driving System (ADS):*** The hardware and software that are collectively capable of performing the entire dynamic driving task (DDT) on a sustained basis, regardless of whether it is limited to a specific operational design domain (ODD); this term is explicitly used to describe a Level 3, 4, or 5 driving automation system.

- **Advanced Driver Assistance Systems (ADAS):** A broad range of features, including those that provide warnings and/or momentary intervention, such as forward collision warning (FCW) systems, lane keeping assistance (LKA) systems, and automatic emergency braking (AEB) systems, as well as some convenience features that involve Level 1 driver support features, such as adaptive cruise control (ACC) and certain parking assistance features.
- **(Human) User:** A general term referencing the human role in driving automation (e.g., driver, passenger, DDT fallback-ready user, driverless operation dispatcher, and remote assistant). These human categories define roles that do not overlap and may be performed in varying sequences during a given trip.
- **(Human) Driver:** A user who performs in real time part or all of the DDT and/or DDT fallback (steering, braking, and acceleration during certain dynamic test maneuvers) for a particular vehicle.



**Figure 1.** The SAE Levels (SAE International, 2021a).

The Connected and Automated vehicles (CAVs) (a.k.a. connected and autonomous vehicles and driverless cars) are transformative technology that has great potential for reducing traffic accidents, enhancing quality of life, and improving the efficiency of transportation systems (Elliott et al., 2019). CAV is a combination of Connected and Automated vehicles, as they are explained in Table 1.

**Table 1.** The Difference between CVs and AVs (Federal Highway Administration, 2018).

Connected Vehicles (CVs)	Automated Vehicles (AVs)
CVs change how drivers interact with each other and the transportation system.	AVs change how vehicles and drivers interact with each other and the transportation system.

With CV technologies, vehicles wirelessly communicate with each other and with infrastructure.	Automated systems perform at least one element of operation without driver input.
CVs enable a range of safety, mobility, and environmental functions. Automated CVs can provide information to drivers.	As automation increases, vehicles are increasingly able to perform dynamic driving functions in varied conditions and environments.
CVs obtain information through wireless communications to support safety, mobility, and environmental applications that assist the driver.	AVs can take over some levels of driving tasks. They can use information through communication technologies to enhance the Automated Driving System capabilities to safely and efficiently interact with the roadway environment.

## 1.2 Literature Review on Cognitive Modeling in Cybersecurity

For the general population, cybersecurity often feels less tangible (Michalec et al., 2023), requiring greater domain knowledge (Khadka & Ullah, 2025), heightened awareness (Ferguson-Walter et al., 2023), and engaging complex multi-stakeholder interactions (Hausken et al., 2024). Given these constraints, humans are regarded as the vulnerable targets. In response, a growing body of cognitive research has focused on modeling human behaviors and cognitive processes in cybersecurity (Maalem Lahcen et al., 2020; Veksler et al., 2018).

Studying attackers' cognitive processes through modelling is considered an effective path to developing mitigation strategies (Meyers et al., 2009) and allocating security resources (Khan et al., 2025). An *attacker* is a person or process that attempts to access data, functions, or other restricted areas of the system without authorization, potentially with malicious intent. Cognitive frameworks have been applied to model attackers' motivations (Dominic et al., 2016; Dutt et al., 2013; Thackray et al., 2016), behavior patterns, and assessing attack impacts (Dutt et al., 2016). These frameworks consider factors from the attacker's cultural characteristics (Sample et al., 2018), risk-adverseness, experience level, and adversarial reasoning. For instance, Koehler and Harvey (2008) used behavioral game theory for attacker's strategies predictions, and Abbasi et al. (2015) employed subjective utilities and prospect theory for attacker's behavior analysis, Veksler et al. (2018) have implemented model tracing through repeated game scenarios to update attacker subjective utilities dynamically. However, as the attacker's decision-making modeling relies heavily on the defense tool's detection mechanism and the defense agent's strategies, this type of models provides a limited view of factors influencing the human behavior and is usually outdated for new threats but gives the human a false sense of system security.

Meanwhile, human performance in cybersecurity constitutes both a potential vulnerability and a critical defensive capability. Defender expertise development is another popular topic in cybersecurity (Aggarwal et al., 2022; Champion et al., 2014; Van Der Kleij et al., 2022). There are some suggestions that traditional approaches to training are less effective when training defensive cybersecurity strategies. For example, classroom-based formal education was compared with informal (self-taught) learning through a Cognitive Task Analysis (Champion et al., 2014). These results indicated that traditional education may not be as valuable as accumulating practical experience. One of the potential reasons is “the breadth of the information available” within the cybersecurity field. Another finding is that cybersecurity practitioners value empirical experience over standard training (Van Der Kleij et al., 2022). Training by simulation games and practice tends to be simplified and delayed. These scenario-based simulations are based on lessons learned from past incidents (Van Der Kleij et al., 2022) and thus cannot cover every possibility. Instruction materials, such as manual books, playbooks, and pre-defined critical incident plans (CIPs), have also been criticized for their rule-based approach and for discouraging deep reasoning (Van Der Kleij et al., 2022).

As a result, training and learning are now often integrated with or replaced by real-time support tools. Some researchers explored how defense-aided tools could aid human self-learning (Andrade & Yoo, 2019), facilitate defenders’ strategies (Abbasi et al., 2015; Dutt et al., 2016), and advance multi-agent collaboration development (Veksler et al., 2018). An autonomous cyber defense team (Prebot et al., 2022) employed the Instance-Based Learning (IBL) theory with the memory dynamics of ACT-R (Gonzalez et al., 2013) to explain the situation awareness of cyber analysts. The integrated model tries to predict the role of intrusion-detection systems in terms of the defender's memory and tolerance for cyber-attack detection (Dutt et al., 2016). The key limitation is still that these tools and models rely heavily on past threats and defense experiences. Therefore, the dataset and experiences, serving the model construction and tool development, are still within a limited scope from the defender’s experiences. In other words, these tools are unable to proactively aid human decision-making with unanticipated incidents from constantly evolving attack types and tactics. Additionally, these aiding tools may increase the complexity of the human operator's task by information overloading, difficulty in interpreting and collaborating, and inducing over-trust.

Another research direction is on assessing and improving the cybersecurity situations of defenders and potential victims (Prebot et al., 2022). The term Cybersecurity Situation Awareness (CSA) has been defined by various research, within different domains, and with various operator

roles (Petersen et al., 2020). Individual factors such as characteristics (Hadlington, 2017), genders (Anwar et al., 2017), past experiences (Hadlington & Murphy, 2018), education levels, and risk preferences (Dutt et al., 2013; Torten et al., 2018) influence CSA and have been broadly investigated. These studies primarily emphasized individual factors and experiences but delved less into how specific tasks and the work environment interact with these factors to influence CSA. Some research (Roy et al., 2010; Sawyer & Hancock, 2018) shifted the focus from human operators themselves to their tasks, showing that the frequency, recency, and patterns of attacks impact CSA.

Gutzwiller et al. (2020) referred to one of the main challenges in cognitive analysis in cybersecurity, such as CSA measurement, as being closely tied to domain-specific contexts and goals. As cybersecurity always functions as a supporting layer within most work domains, a specific domain-specific analysis is needed to support the construction of cybersecurity goals. Van Der Kleij et al. (2022) also referred to the most pressing problem as balancing the broader operational context in mind while adequately investigating a cyberattack. Namely, in healthcare cybersecurity, cognitive modeling must prioritize operational continuity for human safety, with technological advances in monitoring and diagnosing beyond clinical settings (Coventry & Branley, 2018). In the financial sector, the high stakes of large institutions and the pressures of stringent regulatory scrutiny must be considered (Van Der Kleij et al., 2022). In the energy domain, the decentralization of power generation, the gaps between cybersecurity technology upgrades and the multi-decade lifespan of physical equipment, and the requirement for resilient response controls to mitigate cascading safety-critical impacts are critical for cybersecurity (Krause et al., 2021). In the CAV domain, domain-specific considerations include drivers' diminished engagement in the driving task, overreliance and overtrust in automated driving systems, and the time-critical nature of driving (Linkov et al., 2019). Collectively, considering solely general attack patterns or individual human characteristics would be insufficient to accurately reflect human capabilities in response to cyber threats within a specific work domain context.

### **1.2.1 Task Characteristics in Cybersecurity for Cognitive Modeling**

Based on the findings summarized above, the following task characteristics emerge as particularly relevant to cognitive modeling in cybersecurity applications:

- Cybersecurity tasks are highly interactive among multiple roles as attackers, defenders, and end-users with varying levels of expertise. These roles often have conflicting

interests and compete for limited resources, which directly shape these interactions and decision-making.

- Cybersecurity defense functions as a supportive system operating under both hard technical constraints of the work domain and soft organizational constraints. Human performance in the cybersecurity context is influenced not only by work domain-specific system architecture and technical advancement but also by cybersecurity policies, regulations, and organizational requirements that must be followed to ensure compliant and effective operations.
- At the behavioral level, human decision-making is intense, occurring within an information-rich environment that demands rapid responses under time pressure. These challenges are further intensified by gaps in domain-specific and cybersecurity knowledge, which affect how humans interpret information and execute actions.
- Human responses are highly dynamic across different forms of cyberattacks, whether disguised and stealthy or explicit and distracting. These varying threat profiles trigger different cognitive pathways and response patterns, from design-supported reasoning processes to direct perceptual actions.
- Human decision-making in cybersecurity contexts may also exhibit cascading effects, where an individual's response triggers subsequent reactions across the system. These effects are consistent with the first two characteristics of cybersecurity scenarios: namely, the highly interactive nature and system-level interdependencies.

Overall, these characteristics converge on two primary challenges for cognitive analysis in cybersecurity applications:

First, cyberspace is a highly dynamic environment that constantly evolves and engages multiple interacting roles. Thus, cognitive modeling solely relies on past incidents and strategies may generate outcomes that lag behind current domain developments, introduce excessive information, and even foster a false sense of security. Instead, an effective cognitive model should allow for adaptation to emerging incidents and strategies and can account for expanding roles and shifting perspectives, making it more applicable to predict human performance in the dynamic cybersecurity landscape.

Second, the analysis of the specific work environment in which humans operate is essential. As a supporting layer across various domains, cybersecurity must align with the distinctive constraints of work domain environments, including their technological frameworks, organizational structures, and domain-specific goals. Human response strategies to cybersecurity threats and incident management vary widely across domains, as domain constraints and task objectives differ. Applying cognitive models without incorporating adequate domain-specific considerations can result in outcomes that lack real-world transferability. This challenge thus lies in systematically and effectively integrating domain-specific knowledge and contextual factors into the cybersecurity cognitive modeling.

These two main challenges motivate the introduction of Cognitive Work Analysis (CWA) and ACT-R (Adaptive Control of Thought - Rational). CWA is particularly suited for analyzing complex socio-technical work environments that involve both hard constraints (e.g., physical and technical limits) and soft constraints (e.g., organizational factors, regulations, and policies), and interactions among diverse roles. ACT-R, on the other hand, is effective for precisely capturing human performance in time-critical conditions that require both rapid reactions and information processing across dynamic task environments.

In the following chapter, we will introduce the two modeling approaches in greater detail, given their broad application across domains, theoretical and empirical foundations, and their practical extensions and challenges that support cybersecurity-related modeling.

## Chapter 2 Cognitive Models: Challenges in Modeling Human Performance in Cybersecurity

The term *Cognitive Model* is widely used across fields of human factors, engineering, system design, biology, psychology, and artificial intelligence. In short, a cognitive model is an abstraction of human cognition (Pan et al., 2017). From a systems engineering perspective, a cognitive model is defined as "human psychological and thinking processes before the execution of a certain action" (B. Wu et al., 2022), or, more concisely, as the modeling of "human cognitive behavior" (Ritter, 2019). In the field of human-computer interaction, Gersh et al. (2005) defines the cognitive model as "applying cognitive science principles to the design and analysis of complex systems, focusing on optimizing human-system interaction." Collectively, in a broad definition, cognitive modelling refers to both the analysis and computational simulation of human problem-solving and decision-making processes.

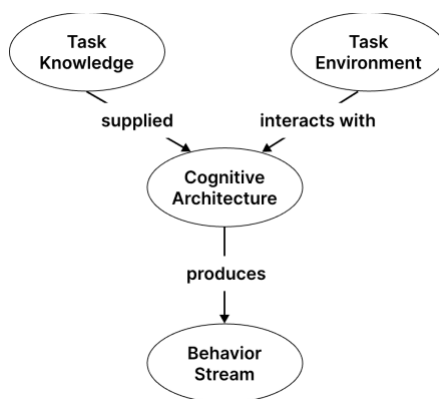
Within the field of cognitive engineering, a *cognitive model* is the process of translating cognitive theories and architectures into engineering applications for quantitative measures of human performance. John (1998) defines it as "computational representations that simulate human cognitive processes to predict user behavior and inform interface design". The computational representations of a cognitive model refer to the description of aspects or functions of human cognition through representations, mechanisms, and processes that can be simulated (McClelland, 2009; R. Sun, 2008). In that sense, AI computing algorithms simulate how humans process thoughts and perceptions are also referred to as "Cognitive AI Models" (Garza-Ulloa, 2022).

Building on these general definitions of cognitive models, this chapter will further review the commonly used cognitive computational models in detail, examining their foundations, focuses, strengths, and challenges in simulating human performance in cybersecurity. Finally, we will explore a brief example to elaborate on the practical challenges of a frequently used cognitive model as applied in cybersecurity human performance modeling.

### 2.1 Cognitive Architectures and Models

A *cognitive architecture* is defined as a system capable of producing all aspects of behavior, while remaining constant across various domains and knowledge bases (Anderson et al., 2004; Newell, 1994). Briefly, a cognitive architecture specifies the underlying infrastructure for an intelligent modelling system that could be applied in various domains (Langley et al., 2009). A similar definition

on cognitive architectures’ consistencies in producing different behaviors with common mechanism is with Langley (2006) “(The architecture) are often used to explain a wide range of human behavior, and to mimic the broad capabilities of human intelligence”, and with Lehman et al. (1996): “a cognitive architecture is two things at once. First, it is a fixed set of mechanisms and structures that process content to produce behavior. At the same time, however, it is a theory, or point of view, about what cognitive behaviors have in common.” Byrne (2012) defined a cognitive architecture as an attempt to build an integrated theory that encompasses a broad spectrum of what is known about human cognition and performance. Byrne (2012) also emphasized that a cognitive architecture alone is generally not able to describe human performance on any particular task; it must be given knowledge about how to do the task. Another way of defining cognitive architectures is a proposal about the mental representations and computational procedures that operate on these representations, enabling a range of intelligent behaviors (Kotseruba & Tsotsos, 2016).



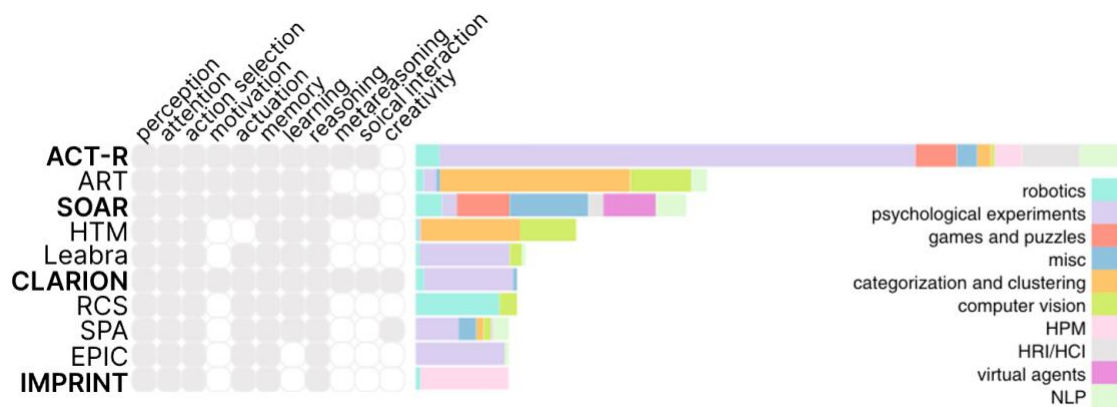
**Figure 2.** Structure of a Model based on a Cognitive Architecture (Byrne, 2012).

In this sense, a *model* of a task based on a cognitive architecture (generally termed a “*cognitive model*” in cognitive engineering) consists of the architecture and the requisite knowledge to perform the specified task (Byrne, 2012). That is, a cognitive model generally comprises three components: the architecture, task knowledge, and a dynamic task environment with which the model interacts (Byrne, 2012). The output of this model is a behavior stream, as in Figure 2. Thus, the cognitive models, beyond the theories and architectures grounded in cognitive science, offer practical value in simulating human performance in different domains and task environments. Specifically, the model can generate performance metrics such as execution time and accuracy. It can also help explain

the underlying sources of human cognitive processes that contribute to these performance differences (Kieras et al., 2001), from the cognitive architecture on which it is based (Ludwig, 2005).

### 2.1.1 Review of Cognitive Modeling: Applications with Different Emphases

Cognitive models are the computational applications of cognitive architecture (Kotseruba & Tsotsos, 2016, 2020). With the large number of cognitive architectures (55 are still actively developed (Kotseruba & Tsotsos, 2016)), most cognitive engineering research focuses on a few models. Arguably, some of these architecture-based models may seem more widely used than others and attract more attention. Figure 3 is from a review paper on the development of current cognitive architectures and their applications across various domains. Among them, ACT-R stands out as the most widely applied, historically significant, and multi-dimensionally competent architecture (Adams et al., 2012; Kotseruba & Tsotsos, 2016).



**Figure 3.** A part of the graphical representation of the practical applications of the cognitive architectures and corresponding competency areas defined in (Adams et al., 2012). The stacked bar plots show the practical applications: the length of the bar represents the total number of the practical applications found in the publications and colored segments within each bar represent different categories (see legend) and their relative importance (calculated as a proportion of the total number of applications implementing these categories) (Kotseruba & Tsotsos, 2020).

There are various ways to evaluate and compare cognitive architectures and their practical applications as models (Chong et al., 2007). The most used evaluation methods are based on the model's symbolic structures (Kotseruba & Tsotsos, 2016, 2020) (symbolic systems are typically implemented as collections of if-then rules (Ludwig, 2005; Pew et al., 1998)), psychological validity

(Byrne & Gray, 2003), and different emphasis on functional cognitive modules (e.g., action, learning, reasoning, etc.) (Chong et al., 2007).

The table below (Table 2) compares several widely referenced architectures as in Figure 3: the most commonly used (ACT-R), the most fully developed in terms of competence (CLARION), an architecture frequently compared with ACT-R (Laird, 2021), and the simulation model with the most applications in Human Performance Modeling (HPM) (IMPRINT).

**Table 2.** Comparison of Leading Cognitive Architecture and Related Competency Areas.

	ACT-R (Anderson et al., 2004)	CLARION (R. Sun et al., 2001)	Soar (Laird & Congdon, 2004)	IMPRINT (Mitchell, 2003)
<b>Relationship to Neurobiology</b> (tie to human brain areas)	Yes	Yes	Not directly	No
<b>Psychologically Validated</b>	Yes	Yes	Partially	No
<b>Modules</b>	Perceptual, motor, goal, and declarative memory	Perceptual, motor, goal (declarative & procedural) memory	Perceptual, working memory, long-term memory, and motor	Not a cognitive architecture but a simulation tool with a task network, operator model, and timing parameters
<b>Learning</b>	Production utility adjustment.	Automatic conversion of sub-symbolic patterns into production rules.	Learning (i.e., chunking) occurs when a resolution is found when reaching an impasse.	Not Supported
<b>Symbolic/Emergent/Hybrid</b>	Hybrid	Hybrid	Symbolic	Symbolic

ACT-R is primarily different from SOAR in that it places a strong emphasis on psychological validation. In contrast, SOAR is rooted in a unified theory of cognition, aiming to support general cognitive functioning across domains without task specificity (Chong et al., 2007). As a result, ACT-R can generate detailed behavioral traces, while Soar focuses more on general reasoning processes and planning changes.

IMPRINT, although it is the most frequently used model in HPM (see Figure 3), belongs to a different cognitive modelling approach. It is not based on a cognitive architecture, but instead functions as a simulation toolset that relies on predefined task descriptions and parameters (Mitchell,

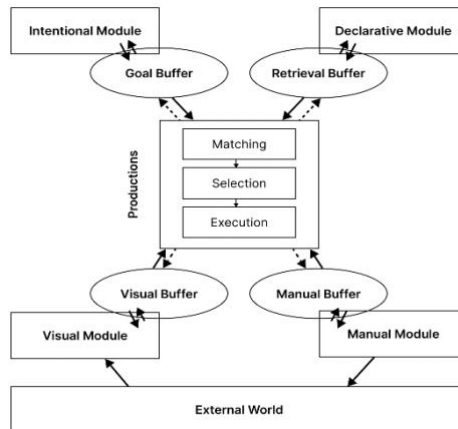
2003). It does not dynamically generate behavior through internal cognitive mechanisms. Therefore, it is better suited for producing estimates of routine-based task requirements, rather than simulating human cognitive variation.

Notably, there seem to be many similarities between CLARION and ACT-R in terms of structure and psychological validity. However, CLARION stands out for its learning mechanisms, particularly in generating new rules through emergent (i.e., ‘creativity’ in Figure 3, which is defined as the computational generation of results that are both novel and appropriate (Kotseruba & Tsotsos, 2016; Still, 2016)), making it a truly hybrid architecture. In contrast, ACT-R’s learning is largely limited to adjusting the *utility* of existing rules, meaning its hybrid nature is more constrained in terms of emergent learning. However, as CLARION emphasizes learning mechanisms, its motor and perceptual timing modules appear less effective than those in ACT-R, particularly in domains requiring precise time-based performance. Compared to ACT-R, CLARION has received less empirical validation and has fewer applications (see Figure 3), resulting in limited tools and extended models.

Therefore, given ACT-R’s strong empirical validation, dominant use in cognitive modeling applications (see Figure 3), and its sufficient ability to produce time-sensitive, detailed behavioral predictions, we primarily consider it as a powerful candidate to model the complex system’s cybersecurity human performance in this work. In the next section, we will examine its core structure in more detail, along with its current applications and development in the context of driving tasks.

## **2.2 Adaptive Control of Thought-Rational (ACT-R)**

The basic architecture of ACT-R consists of a set of *modules*, as shown in Figure 4. Each module is devoted to processing a different kind of information (Anderson et al., 2004). Perceptual modules (e.g., visual module) are for identifying perceived objects (e.g., an object in the visual field), a manual module is for hand-action controlling, the declarative module is for retrieving information from memory, and the goal module is for keeping track of current goals and intentions (Anderson et al., 2004). The modules communicate through their buffers. A *buffer* relays the request to its model to perform actions and respond to queries about the module/itself status.



**Figure 4.** The Basic ACT-R Architecture.

In ACT-R, *Declarative Knowledge* reflects the factual information that a person knows and can report (Anderson & Schunn, 2000). A declarative unit of knowledge is represented in a structure called *chunks*, and procedural knowledge is represented in a structure called *productions*. Each buffer could hold one chunk at a time due to human cognitive capacity limits.

A *production* is a statement of a particular contingency that controls behavior (Bothell, 2022). A production rule defines a set of conditions and actions, such that the conditions must be met for the rule to execute or fire the given actions (Salvucci & Taatgen, 2008). As shown in Figure 4, the solid arrows show which module reads the information from another module's buffer, and the dashed line arrows show which modules make requests to another module's buffer or directly modify the chunk it contains. This is how a production works in a condition-action manner. This production rule element forms the symbolic core of the cognitive architecture and its models, enabling strong interpretability by allowing researchers to trace and explain the model's predicted human behaviors.

The chunks and productions are all separate pieces of knowledge (Anderson & Schunn, 2000). For connecting them, ACT-R imposes a strong organizational structure through its goal stack (i.e., Intentional Modules in Figure 4), allowing subgoals to be added dynamically (i.e., removed and added by production rules selections), guiding the order in which knowledge is accessed and rules are applied.

Researches also tied ACT-R's basic modules to specific brain regions during complex information-processing tasks by fMRI technology (Anderson, 2007; Anderson et al., 2008), and the

basic modules could predict BOLD responses in specific brain regions (Perceptual modules align with the auditory cortex and fusiform gyrus; processing modules (retrieval and representational change) map to the parietal and prefrontal cortices; and cognitive control is associated with the anterior cingulate cortex and caudate (Anderson, 2007). These biological grounding also supports the plausibility of ACT-R as a symbolic–sub-symbolic integrated system (Kelley, 2003).

Up to this point, we have introduced ACT-R at the symbolic level (Anderson & Schunn, 2000). Next, we will briefly explore ACT-R’s subsymbolic structure and how it works together with the symbolic structure to support the model’s learning functions.

### **2.2.1 ACT-R: Strategies by Symbolic Construction and Sub-symbolic Selection**

The symbolic part consists of declarative and procedural knowledge that is strongly connected to the domain analysis (facts, rules, and physical environments). It is straightforward: different strategies are defined by different productions (or sets of productions) that describe paths to achieve task goals. The subsymbolic part is about processing these components (Anderson & Schunn, 2000). To be specific, when given a goal to achieve (i.e., a problem to solve), the subsymbolic part selects among competing production rules and then executes the selected (Anderson et al., 2004).

In particular, this strategy selection process has two subsymbolic parameters: expected effort and the probability of success. The expected effort is the amount of effort required to obtain a solution if the given production is selected. The expected success rate is the probability that a solution will be obtained if the given production is selected. ACT-R selects among productions by computing the tradeoff between the two to select the production with the best tradeoff value (Anderson et al., 2004). This calculation for strategy selection is of the core subsymbolic mechanism (i.e., “adaptive control” in ACT-R).

The learning of utilities is controlled by the following equation for a production  $i$  after its  $n^{th}$  usage:

$$U_i(n) = U_i(n-1) + \alpha \cdot [R_i(n) - U_i(n-1)]$$

Where  $\alpha$  is the learning rate,  $R_i(n)$  is the effective reward value given to production  $i$  for its  $n$ th usage, and  $U_i(0)$  is the initial utility value for the rule (default to be 0). This calculation of assigning utility values to specific rules serves as ACT-R’s efficient mechanism for rule selection and switching.

By running this symbolic mechanism, the increase in utility from successfully firing specific rules can lead to production compilation, as higher-utility shortcuts emerge and replace more extended sequences involving memory retrieval. This short-cutting reflects skill acquisition, where more efficient strategies are formed through the selection of shorter, high-utility paths. This adaptive compilation allows a problem to be solved with fewer productions and, therefore, performed faster (Bothell, 2022). And this process is how ACT-R interprets the learning process as a gradual translation of declarative to procedural skills (Salvucci & Taatgen, 2008).

Interestingly, this utility-driven compilation process conforms to how rule-based decision-making transformed into a skill-based one, as explained in Rasmussen's Skill-Rule-Knowledge (S-R-K) framework (Rasmussen, 1983). The details of how these two cognitive frameworks conform in understanding skills development will be mentioned in the later sections (see Section 4.3.3).

However, the limitation of this subsymbolic mechanism is the limited 'creativity' for emergent strategies (see Figure 3). This limitation has already been noted and recognized from ACT-R's inability to develop learned recognition, a feature typical of true subsymbolic systems (Kelley, 2003). Building on ACT-R's established architecture and mechanisms, its demonstrated applicability and ongoing developments in modeling human driving performance across varying conditions and behavioral metrics. And we will use an extended ACT-R (QN-ACTR) model for an attempt at simulating human responses to distraction-based cyberattacks on CAV systems in Chapter 3.

### **2.3 Research Questions**

While cognitive models offers a promising approach for identifying and mitigating human vulnerabilities, it continues to face additional challenges in achieving practical applicability to the complex, dynamic in cybersecurity with sufficient domain-specific considerations. Given these intertwined challenges, a robust cognitive modeling approach for cybersecurity must effectively integrate both tailored domain analysis and solid simulation of human dynamic behaviors within a complex environment.

To meet these expectations, Cognitive Work Analysis (CWA) (Rasmussen et al., 1994; Rasmussen & Jensen, 1974; Vicente, 1999) and Adaptive Control of Thought-Rational (ACT-R) (Anderson et al., 2004) stand out. Both are classic and well-established, consistently contributing to the development of cognitive modelling and analysis, but with different emphases.

Cognitive Work Analysis (CWA) (Rasmussen et al., 1994; Rasmussen & Jensen, 1974; Vicente, 1999) is a formative cognitive analysis framework that supports systematic domain-specific analysis for complex sociotechnical systems. It thus can enable a comprehensive and multifaceted understanding of cybersecurity within its specific work domain.

Adaptive Control of Thought-Rational (ACT-R) (Anderson et al., 2004), on the other hand, provides a mature foundational cognitive architecture for computational models to dynamically simulate human performance in various task environments.

Combining the strengths of both, we propose an integrated cognitive model that incorporates CWA's domain-specific analysis into an ACT-R-based computational model's simulation of human performance, addressing the challenges in applying cognitive modeling to cybersecurity in complex domains.

Before integrating the model, we need first to examine the standalone applicability of the ACT-R model in cybersecurity and identify the necessary enhancements, by using CAV cybersecurity as a representative application domain in the next chapter.

**Research Question 1:** *Is an ACT-R model sufficient for modeling human performance in cybersecurity?*

To answer this question, Chapter 2 will review commonly used cognitive computational models and their core architectures. This review will further explain the rationale for selecting the ACT-R model and discuss its potential limitations when solely applied to cybersecurity in a specific domain. Chapter 3 will illustrate an example of extended ACT-R model, suggesting both the benefits and challenges of applying an ACT-R model in this context.

Based on Chapter 3's findings, we will then propose an enhanced model that integrates the CWA insights to address the limitations of using the ACT-R model alone. Chapter 4 will elaborate on the conceptual integration of CWA insights into the construction of an ACT-R model for a complex work domain, continually using CAV cybersecurity as an illustrative example. The chapter will provide a detailed exploration of each CWA dimensions' integration into the ACT-R model and answer the second research question:

**Research Question 2:** *Can insights from Cognitive Work Analysis (CWA) enhance the effectiveness of an ACT-R cognitive model in complex domains such as cybersecurity?*

To validate the potential of integrating CWA's domain-specific insights with the ACT-R model, it is essential to identify a suitable use case that engages effective cybersecurity risk management to demonstrate the model's enhancements.

Hence, Chapter 5 will continue using CAV cybersecurity as the illustrative example to assess the effectiveness of in-vehicle human responses to cyberattacks. The assessment will investigate the in-vehicle human analytic process and response patterns to different attack forms, and the factors shaping these behaviors, to determine if it's an appropriate task environment to simulate effective human performance to cyber threats, as stated in the following research question:

**Research Question 2a:** *Is in-vehicle operation an effective task environment for detecting and mitigating cybersecurity threats?*

Following the question of identifying an effective cyber threat management task environment as the model's use case, the development of Cybersecurity Operations Centers (SOCs) across various domains serves as a supportive sector within the general cyber defense framework, providing another option for applying the cognitive model to simulate human performance in effectively detecting and mitigating cyber threats. Similarly, we will assess the effectiveness of the SOC work environment for cybersecurity management and the application of the cognitive model:

**Research Question 2b:** *Is a Security Operations Center an effective task environment for detecting and mitigating cybersecurity threats?*

Compared to in-vehicle operations, SOC analysts are expected to perform more domain-specific analytical decision-making for cybersecurity management. Chapter 6 accordingly develops and applies the integrated cognitive model to simulate SOC's human analyst task performance. To validate the improvements of the cognitive model, its simulation results will be compared between the integrated model and a baseline model. Through model construction and result comparison, Study 2 aims to answer the third research question, which focuses on the implementation and development of the cognitive model:

**Research Question 3:** *What are the enhancements, limitations, and future directions of integrating CWA and ACT-R for modeling human performance in complex domains like cybersecurity?*

Frontier work (Collins et al., 2022; Joshi & Ustun, 2024) has explored the integration of Generative AI (GAI) models with classical cognitive models to simulate human-like decision-making within complex contexts. GAI models (Malloy & Gonzalez, 2024), although suffering from many limitations due to their inherent differences from classic cognitive models, are nonetheless advancing the development of cognitive models (Malloy & Gonzalez, 2024). Therefore, a follow-up study will also test our enhanced cognitive model against a representative GAI model to examine the opportunities and limitations of future cognitive modeling development, in Appendix C.

## 2.4 Thesis Structure

The first half of this thesis focuses on developing an effective cognitive model for simulating human performance in complex domains, using CAV cybersecurity as an illustrative example. This part begins by reviewing current prevalent cognitive models, with their foundational architecture and distinctive strengths and limitations. It concludes by finalizing the selection of ACT-R models as the base model. We then identify the gaps in applying the ACT-R model alone in the CAV cybersecurity domain (Chapter 3) and accordingly propose an enhanced solution by integrating CWA insights into the model (Chapter 4). To validate the potential enhancements, Study 1 examines human responses to cybersecurity threats during in-vehicle operation. However, we learned that human drivers' in-vehicle responses to cyber threats are relatively simple and don't involve deep domain-driven decision-making. Therefore, while a CWA-enhanced model may offer improvements, the value of a CWA-enhanced ACT-R model cannot be well-explored in this domain.

We then decided to explore a cybersecurity operation center environment to demonstrate the model's enhancement in a context with more complex decision-making (Study 2). We begin by demonstrating the development of the integrated model in this environment, using this as an opportunity to examine the model development and explore empirical validation of the model. A discussion of the integrated model's application to other domains, its limitations, and future improvements in light of rapidly developing AI models, concludes the second half of the work.

A brief overview of each chapter is presented below:

- *Chapter 2* reviews the prevalent cognitive architectures and models. This chapter explains our selection of ACT-R models for simulating human decision-making in cybersecurity.

- *Chapter 3* illustrates a brief example using an extended ACT-R model with adjustments to simulate in-vehicle human operations when confronting distraction-based attacks targeting CAVs. This example implied limitations to directly applying the ACT-R model alone in cognitive modelling efforts in the cybersecurity context.
- *Chapter 4* suggests integrating CWA's insights into the ACT-R model to narrow the cognitive model's gaps. This chapter provides a further introduction to the CWA across its dimensions. In comparison, CWA, as an analysis framework, provides a valuable structure for guiding domain-specific analysis, whereas the ACT-R model serves as a computational model simulating human behavior. This chapter will then expand on their distinct foundational roles as cognitive analysis approaches, exploring and discussing their compatibility in application, detailing how CWA-guided insights support the construction of the ACT-R model and enhance its applicability to complex systems.
- *Chapter 5* investigated whether in-vehicle operation is an effective cybersecurity management task environment for applying the proposed model integrating CWA and ACT-R. This chapter used a survey study (Study 1) to examine the awareness and response of in-vehicle human drivers to different forms of cyberattacks. The study results, however, reveal that in-vehicle human responses to cyber threats are less involving analytic decision-making, often reverting to a safe state rather than solving or mitigating the cyber threat. As a result, this may not be an effective cybersecurity management task environment for applying the enhanced model's capability in domain-specific analysis processes. This leads to a shift in our modeling focus from direct system end-users, such as in-vehicle drivers, to human roles in the supportive sectors of cybersecurity defense frameworks, namely the Security Operations Center (SOC) analysts.
- *Chapter 6* details the process of integrating CWA's domain-specific insights into the ACT-R-based model construction and validates the model's enhanced effectiveness in SOC analysts' task (Study 2). Study 2 applies the CWA framework step by step to generate a SOC tier-1 analyst's task environment analysis, systematically supporting the construction of an ACT-R model with domain-specificity. We then applied the integrated cognitive model to simulate human performance in the cybersecurity alert triage task. The model's enhancement was validated by comparing it against a baseline model without CWA insights, based on how well

each fit the collected empirical. The integrated model demonstrates a marginal trend toward higher quantitative accuracy in predicting human processing time and final decisions (ACT-R) when informed with sufficient domain-specific analysis (CWA). Its improved estimation provides clear interpretation and measurable validation into analysts' adaptability and vulnerabilities in security alert processing tasks. We then accordingly discussed strengths, limitations, and application considerations of the integrated cognitive model.

- *Chapter 7 summarizes* insights on our cognitive modelling effort and discusses the challenges with future development in applying the CWA-informed ACT-R model to cybersecurity and other complex domains.

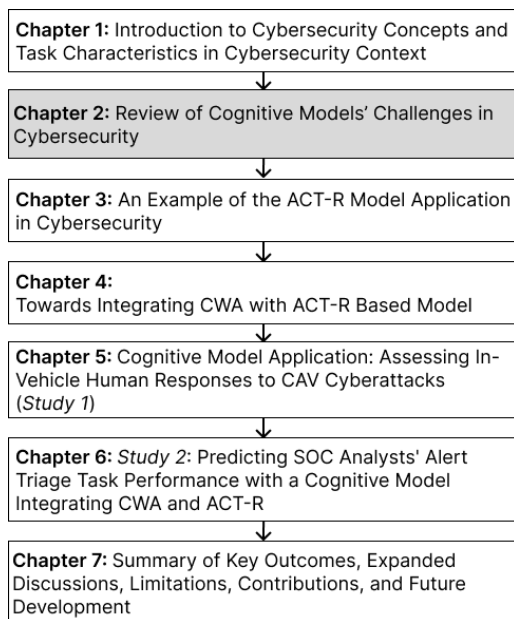


Figure 5. Overview of Thesis Structure.

However, given the integrated model's limited applicability in handling ambiguous conditions and the rapid development of GAI models during this work, we conducted a brief comparison study to evaluate the decision-making prediction accuracy of a large language model (LLM) against it in Appendix C. The findings then lead to an expanded discussion on the strengths and insights that the classic cognitive modeling approaches can offer for the application and development of GAI models in cognitive models' development.

## 2.5 Contributions

**Contribution 1:** This work will be the first attempt to explore the possible advantages of integrating ACT-R and CWA when modeling human performance in complex domains like cybersecurity. The detailed integration process supports the rationale of their compatibility and improvements in model construction efficiency, and the results confirm the enhancement of the output, thus contributing to the development of a more robust cognitive modeling approach that fits a specific, complex sociotechnical working environment, incorporating both specific domain fidelity and general quantifiable measurements.

**Contribution 2:** The cognitive model development process will illustrate a practical application of the CWA analysis to enhance the domain-specificity of a computational model. The pipeline, comprising the identification of the effective task environment, developing the domain knowledge and analysis following CWA framework, integrating these analysis into the ACT-R model's components, and consideration of strength, trade-offs and limitations in modeling strategies, would contribute as a practical reference for guiding future efforts in developing computational models with improved domain specificity by incorporating CWA-informed insights.

**Contribution 3:** This work will extend the cognitive modelling focus from end-users' response to the supportive human roles involved in the broader defense framework. The broader modeling scope and expanded dimensions of human roles may offer deeper insights for improving cybersecurity-aided design and advocating for the development of a collaborative defense framework, particularly in the context of the CAV landscape.

**Contribution 4:** This study will also contribute to clarifying the strengths and challenges of traditional cognitive models and introducing new perspectives on how traditional models can be enhanced and selected based on task contexts, and how AI models' advancement can learn from classic cognitive models to improve interpretability and robustness in domain-specific applications.

## **Chapter 3 An Example of the ACT-R Model Application in Cybersecurity**

We here choose the CAV cybersecurity domain as an example to illustrate domain-specific challenges in cognitive modeling for three reasons. First, CAV is a rapidly evolving technological landscape with growing cybersecurity concerns (X. Sun et al., 2022). The fast advancement of driving systems adds complexity and dynamism to their cybersecurity support operations, posing challenges for human cognitive modeling to adapt to these evolving conditions. Second, the CAV domain exhibits prominent domain specificity (Linkov et al., 2019), characterized by time-critical operations and highly collaborative interactions involving multifaceted stakeholders (e.g., in-vehicle users, external road users, service providers, etc.) and extensive dynamic road conditions. These factors highlight human flexibility and prompt decision-making in cognitive modeling for cybersecurity. Third, cognitive modeling efforts in CAV are limited compared to the extensive technical research in the field (Linkov et al., 2019), reflecting a pattern common across many domains of cybersecurity research.

These factors of CAV cybersecurity well represent the cognitive modeling challenges outlined in Chapter 2. We therefore use CAV cybersecurity, in this chapter, as an example to examine these challenges in detail.

### **3.1 ACT-R Models Strengths in CAV Cybersecurity**

ACT-R models have long been used for simulating human interactions with in-vehicle technology over the past 40 years (Park & Zahabi, 2024).

The basis for its prevalence in driving tasks is that the ACT-R model fits well with the action-intensive and time-critical nature of the driving domain. ACT-R incorporates Fitts' Law into its perceptual-motor system, enabling the precise simulation of humans' perception-to-movement processes (Anderson et al., 2004). The developed ACT-R extended models with different focuses have successfully simulated a wide range of drivers' actions, including braking (Rehman et al., 2024), lane changing, steering, takeover scenarios, training effects, and responses to urgent situations (e.g., maneuvering between a lead and following vehicle) (see (Park & Zahabi, 2024) for a review).

The advancement of automated driving has increased the need for modeling driver distraction and multitasking. But ACT-R is inherently a serial processing architecture with limitations in

dynamically simulating threaded tasks. To address this limitation, extensions such as threaded task processing (Salvucci & Taatgen, 2008) and the parallel processing capabilities from the QN-MHP (Y. Liu et al., 2006) framework have been incorporated into ACT-R models. One notable advancement is the QN-ACTR model (Cao & Liu, 2013b, 2013a, 2014), which integrates features of both ACT-R and Queueing Network models to better simulate parallel perceptual processing for multitasking in driving tasks (Deng, Cao, et al., 2019b; L. Xu et al., 2025).

Collectively, the continuous improvement and widespread application, with empirical validation, provide extensive resources for applying the ACT-R-based model in complex driving contexts. Naturally, CAV cybersecurity serves as our first example application domain for assessing the use of ACT-R models in cybersecurity. In this chapter, we examine the challenges of applying current ACT-R models to cybersecurity, with a specific focus on the domain specificity of the CAV context. We adapted the QN-ACTR model from simulating multitasking in driving takeover performance and augmented it to simulate three distraction-based attacks that interfere with driving tasks (Deng, Cao, et al., 2019a). The primary purpose is to assess potential gaps in an ACT-R based model construction and application for simulating human performance in CAV cybersecurity scenarios, and to explore potential approaches for improvement accordingly.

As this early-stage simulation is to quickly assess the applicability of the ACT-R model in the CAV cybersecurity, we will not empirically evaluate the model results but compare the results to performance impairments associated with different blood alcohol concentration (BAC) levels as a reference metric.

### **3.2 Distraction-Based Cyberattacks Targeting In-Vehicle Human Operations**

Among the various attacks on advanced driving systems, a particularly concerning type is to distract human drivers. These attacks impair human cognitive capacities by interrupting timely control in hazardous situations (R. Sun et al., 2001). Many studies showed that with ADAS' facilitations, drivers often divert their attention from the primary driving task to secondary activities or the in-car environment, increasing the likelihood of exposure to distraction-based cyberattacks (Hungund et al., 2021). Furthermore, ADAS proves to reduce human skills and knowledge of driving (Hadlington & Murphy, 2018; Linkov et al., 2019; Noy et al., 2018). Such trends intensify the risks of distraction-based attacks as drivers' focus can be more easily manipulated by tampered interruptions or alterations of the driving environment. Given that distraction is continuously among the leading

causes of vehicle accidents (Gershon et al., 2019; Strayer et al., 2006), modeling human performance confronting distraction-based attacks is the focus of this chapter.

The BAC level is a widely accepted benchmark for assessing driving performance and cognitive status (Strayer et al., 2006). It has also been effectively used to compare the impact of dual-task activities on driving performance, like cell phone use while driving (Strayer et al., 2006). In assessing driving performance, *reaction time* is among the most widely analyzed behavioral measurements (Jongen et al., 2016; Strayer et al., 2006). As a prior experiment collected human driving performance with different BAC levels, we will use the collected empirical data as reliable benchmarks for assessing cognitive impairment by modelling distraction-based attacks, gauged through delayed reaction time (Christoforou et al., 2013; Yadav & Velaga, 2019).

### **3.2.1 Assumptions**

Based on the current review and understanding of ACT-R model, we have two assumptions for this model example.

First, while ACT-R models are effective for simulating driving tasks, their suitability for modeling human responses to CAV cyber threats remains uncertain. ACT-R is good at modeling tasks involving intensive perceptual-motor control with routine procedures (N. Taatgen & Anderson, 2010). However, handling cyber threats necessitates more analytical information processing and exploratory diagnostic (D'Amico et al., 2005) than regular driving tasks. Given that, the model's construction of production rules for the analytic process of complex cyberattacks would substantially increase modelling effort. Still, it may not fully demonstrate the model's strengths in perceptual-motor simulation. Based on this assumption, we expected the ACT-R-based model to be more effective at capturing human responses to perceptually salient cyberattacks. As a starting point, our first modeling attempt begins with simulating distraction-based attacks.

Another consideration is that the model's frequently used subtasks for simulating multitasking and distraction are traditional psychological tasks with abstract content (i.e., these tasks are designed to represent generalized cognitive processes rather than domain-specific activities). To better approximate distraction-based attack scenarios, we assume that these general subtasks need to be augmented with domain specificity to increase the fidelity of the scenario representing the attacks. However, as this modeling effort is still in its early phase, with limited data available on in-vehicle cyberattacks, we will still use the general psychological subtasks with minor adjustments to

approximate distraction-based attacks. Yet, we acknowledge these limitations in reflecting the domain-specific fidelity of CAV cyberattacks and will discuss the effects of this setup on the modelling process and results in a later section (see Section 3.3.2).

### 3.2.2 Model Construction

The foundational model we used is QN-ACTR. It is an ACT-R extended model integrated with the queueing network to handle multitasking modeling (Cao et al., 2014, 2015; Cao & Liu, 2015). To adjust the model, we attempted to map the collected participants' feedback (see Appendix A for the Informal Interviews **for Collecting Participants Insights**) to the production rules outlined in (Deng, Cao, et al., 2019a) to augment the model's subtasks representing distraction-based attacks.

The design of the three scenarios was primarily guided by the technological feasibility of cyberattacks and the foundation model's ability to represent them through visual and auditory stimuli. We focused on multimodal distractions, as recent research suggests they pose greater risks than single-modality distraction attacks in driving contexts (Q. Zhong et al., 2024). In addition, we included scenarios that extend beyond distraction alone and require drivers to engage in diagnostic reasoning based on anomalies or misinformation, as the analytical processes are essential for effective cyber threat detection and mitigation (Colabianchi et al., 2025). The simulated attack scenarios, therefore, combine visual and auditory stimuli as distraction-based attacks, along with scenarios that additionally require diagnostic reasoning. The simulated attack scenarios are as follows:

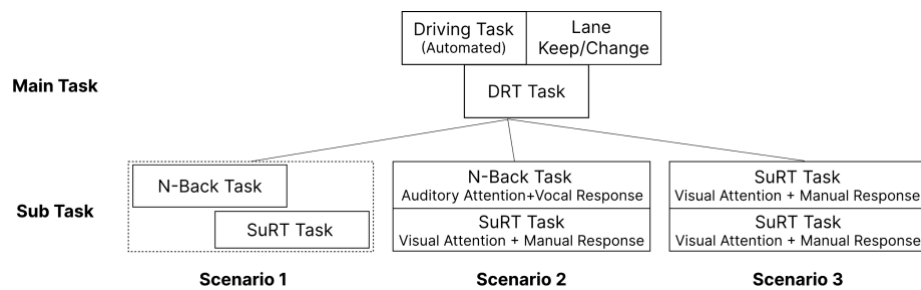
- Concurrent Auditory Visual Distractions (*Scenario 1*): This scenario features simultaneous auditory and visual distractions. It doesn't necessitate any additional anomaly diagnosis by the driver. An example of this modeled scenario involves the driver interacting with a tampered in-vehicle infotainment system that suddenly increases the music volume. Simultaneously, a non-urgent indicator light appears on the dashboard, prompting the driver to turn it off.
- Auditory Intervention and Visual Inspection (*Scenario 2*): This is an aural stimulus attack, prompting a human driver's visual inspection. This scenario could represent misleading navigation instructions from a voice assistant, triggering functions without the driver's consent. Such stimuli demand a human driver's interpretation and elicit vocal responses to cancel a tampered input. Drivers also need to visually verify the status of engaged systems, such as ensuring a voice assistant's navigation information matches the GPS display.

- Visual Intervention and Inspection (**Scenario 3**): This is a scenario where human drivers process visual interventions with visual inspection from another information source. When presented with false visual notifications (e.g., false visual alarms from a collision detection system), a human operator may need to refer to the dashboard for other information sources to confirm the system’s actual condition.

The QN-ACTR model for the driving multitasking simulation (Deng, Cao, et al., 2019a) comprises five primitive task models: (1) a driving task, (2) the SuRT task, (3) an N-back task, (4) a Detection Reaction Task (DRT), and (5) Lane Keep/Changing reaction. The Surrogate Reference Task (SurT) and N-back tasks are standardized general psychological tasks commonly used to model visual search with manual responses and auditory attention with vocal responses, respectively.

In our implementation, we retained the driving, Lane Keep/Changing, and DRT tasks together, simulating driving performance to get reaction time. In contrast, the SuRT and N-back tasks were reorganized to represent different distraction-based attack patterns in both visual and auditory modalities. In other words, our adjustment to the foundational model involved reorganizing sub-tasks to represent the Attack Response tasks.

An illustration of the reorganized subtasks is provided (see Figure 6). In Scenario 1, the two sub-tasks are not serially connected, as this attack does not require sequential reactions and analysis. In Scenarios 2 and 3, the two subtasks were connected as the outputs from the first task required the second task's action to align with.



**Figure 6.** The Basic Task Models from (Deng, Cao, et al., 2019a) and the Reorganization of these Task Models for the Modeling Test.

**Main Tasks:** Deng et al. (2019a)’s driving task model is rooted in Salvucci’s driving model (Salvucci, 2009; Salvucci & Taatgen, 2008) and has been further refined by incorporating mirrors as

additional visual zones (Cao & Liu, 2013b, 2015). The DRT task is for handling task switches in takeover requests, and the lane-keeping / -changing tasks simulate drivers' driving actions.

**Sub Tasks:** Prior studies (Cao & Liu, 2015; Deng, Cao, et al., 2019a) used the generalized visual and auditory task models to simulate distracting tasks. These task models cover a range of multi-modality inputs and outputs with visual and auditory signals and manual and vocal responses (e.g., N-Back Task, SuRT).

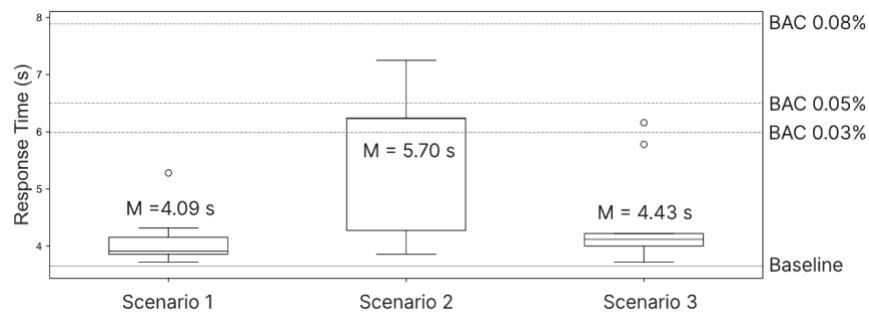
**Reorganization of Subtasks for the Attack Response Model:** Our assumed distraction-based attacks fall into three categories, based on their modality and the drivers' response. Beyond only perceptual distractions, we included scenarios that demand further information inspection, based on participants' feedback, indicating that drivers tend to ignore attacks unless additional analysis is required to ensure driving safety. Namely, we define one modelled scenario (Scenario 1) as purely perceptual, requiring no further information analysis, while the other two (Scenario 2 & 3) involve distractions that necessitate additional inspection and analysis of the distracting information to maintain the driving task.

The differences among the three attack patterns were represented in how subtasks are coordinated and organized (see Figure 6). In scenario 1, attention is drawn to the two different modality subtasks that do not interfere with each other and do not require immediate action. In Scenarios 2 and 3, the subtasks are structured serially: the first subtask generates a response and triggers a follow-up subtask. This serially connected subtasks and the requirement for responses thus compete with driving and DTR tasks for cognitive resources in the model simulation run.

### **3.2.3 Model's Simulation Results and Evaluations**

For each attack scenario, the three augmented models (see Figure 6 for the scenario-specific models with their reorganized subtasks) were each executed ten times to generate scenario-specific simulation results. The parameters and models coefficients remained unchanged from the foundational model. We then compared the simulation results with the empirical data from Yadav and Velaga (2019), who reported reaction times under varying BAC levels in a driving takeover task setting similar to ours (See Appendix A for details on the driving environment setting and evaluation references).

Upon the comparison, the concurrent perceptual distractions (Scenario 1) do not result in reaction times exceeding a BAC level of 0.03% (see Figure 7). This result suggests that even when subjected to multi-modality distractions, if the distraction doesn't necessitate a complex cognitive process from the driver, it only imposes limited impacts on the driver's reaction time. In Scenario 3, where visual anomalies serve as a distraction that necessitates a following information inspection visually, specific outcomes surpass the reaction times observed at a BAC level of 0.03%, indicating that subsequent analysis can elevate the cognitive load on human drivers to a risky level.



**Figure 7.** The Comparisons of Simulated Reaction Time vs. Reference Values.

For Scenario 2, which involves auditory anomalies requesting a subsequent visual inspection, six of the ten simulations resulted in reaction times exceeding that of a BAC level of 0.03%, leading to risky consequences (see Figure 7). In this scenario, an auditory distraction and a follow-up visual inspection amplify the risk, leading to prolonged reaction times compared to the Scenario 2 model. This implies that transitions between perceptual modalities, especially when intertwined with advanced analysis processes, can severely extend driver reaction times, potentially escalating the risk from the attack.

### 3.3 Discussions of Applying ACT-R Model in Cybersecurity

The above applied an ACT-R-based model to run a preliminary simulation of human driver performance with distraction-based attacks.

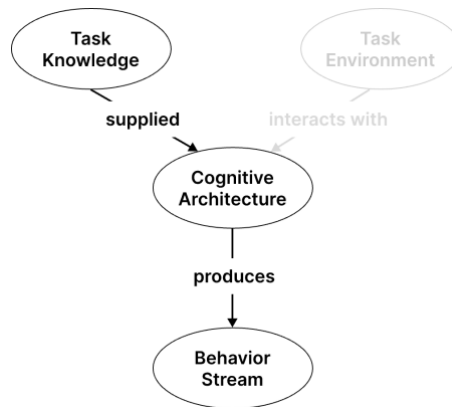
The ACT-R model has demonstrated strong value through its flexible architecture that adapts to task-specific needs and solid empirical validation with human data, allowing for its great potential in simulating human responses under multimodality distraction-based cyberattacks. However, this attempt also reflects some limitations in using the ACT-R model alone for cybersecurity applications.

The limitations are grouped into two categories: the first emerged during the modeling process itself (sections 3.3.1 and 3.3.2), while the second is about the model's application concerns (section 3.3.3).

### **3.3.1 Insufficient Systematic Analysis to Support Rule Construction and Mapping**

Firstly, we observed that the process of constructing the model's production rules tends to be subjective, lacking a systematic analysis framework for guidance. There was no standardized framework to follow for analyzing the participant feedback and translating it to the model's components. The reorganization of the subtasks is based entirely on the modeler's understanding and assumptions, informed by the modeler's own experience and an intuitive understanding of how the subtasks can represent attack patterns. Thus, the simulation results are less convincing in assessing the risks of the distraction-based attacks.

Although our initial exploratory modeling was informal and exploratory, this subjectivity in constructing and mapping production rules appears to be common across a few modeling studies (Dimov et al., 2020; Marewski & Mehlhorn, 2011). The analysis for model rule construction is often based on the modeler's understanding of the task and task-specific analysis, with limited examination of how humans in a real, complex work-domain environment infer and make decisions, considering broader factors systematically. Both the post-hoc discussions from our exploratory modeling and the literature review on rule construction in related work reveal a tendency to build production rules without a broader system-level analysis of the task and validation. This tendency has also been noted to result in models that fit observed data rather than systematically predict behavior (Dimov et al., 2020).



**Figure 8.** Structure of a Model based on a Cognitive Architecture (Byrne, 2012) (focused on Task Knowledge).

This concern about the ACT-R model construction process is not new. Marewski and Mehlhorn (2011), for instance, compared various decision-making assumptions drawn from the literature by implementing them in an ACT-R model for a simple task. When these different assumptions were mapped into the model construction, even for the same task, the simulated results varied significantly. To more accurately reflect human cognitive processes, especially when handling complex information, such as responding to cyber-attacks on vehicles, a formal analysis of task knowledge (see Figure 8) based on real human insights is essential. Strategy selection and rule mapping should follow a validated framework to improve the model’s construction stability and consistency within defined task domains.

One potential reason for many ACT-R models’ heuristic and direct rule construction process is that, as previously mentioned, the model is most effective and prevalent in simulating perceptual-motor tasks following routine task procedures. The driving takeover task is a good example with intensive perceptual-motor components (e.g., visual attention, braking, and steering) with a clear procedural structure (perception leads to coordinated motor actions). In this task condition, the construction of rules can be straightforward, as the task’s procedural knowledge is commonly admitted being routinized skill-based operations.

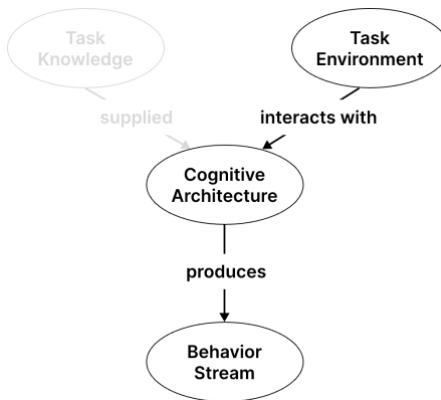
In contrast, constructing rules for tasks that involve both skill-based actions and analytical reasoning is far more demanding. These tasks require broader consideration of domain knowledge and system-level factors, which is typical in cyber-threat response work. The tasks in complex work-domain systems involve both perceptual–motor activities and higher-level reasoning. They rely

heavily on domain-specific declarative knowledge and on system-level constraints beyond the immediate task. These factors guide how rules are matched, selected, and switched in the simulation of human performance. As a result, the construction of models based on work domain knowledge and systematic analysis of work complex environments becomes highly essential for defining rationale and precise firing and transition conditions. With such tasks, relying only on the modeler's subjective interpretation or on task analysis alone is insufficient. Without a structured, system-level work-domain analysis framework, the model becomes vulnerable to bias and instability (Marewski & Mehlhorn, 2011). It also fails to account for the full complexity of the task, which ultimately reduces the reliability of the simulation results

### **3.3.2 Generalized Task Environment**

Second, the model tends to represent task environments in a simplified manner (see Figure 9).

Our modeling results suggest that mere distractions do not excessively strain human drivers' cognitive capacities. In contrast, the risk can increase significantly if the attack requires further driving information inferences by the driver. However, since the cognitive tasks modeled here are abstract and generalized, and the modelled distraction attack content is not drawn from real-world cases, the specific in-vehicle information manipulated or affected is not clearly defined in our model. It is challenging to conduct a detailed analysis of the human driver's strategies in information processing to decision-making with abstract task content, and to give feasible mitigation solutions accordingly. This gap echoes the previously assumed challenge in using an ACT-R model for CAV cybersecurity: its insufficiency in capturing the complexity of cybersecurity-related task environments.



**Figure 9.** Structure of a Model based on a Cognitive Architecture (Byrne, 2012) (focused on Task Environment).

Drivers' decision-making processes and responses vary depending on the specific content of distracting information. In particular, with the advancement of ADAS technologies and the great likelihood of distraction-based attacks targeting these systems, it is essential to identify the exact ADAS information that has been tampered with or manipulated to divert the driver's attention. Since information provided by ADAS varies in modality, content, accessibility, and its criticality to the driving task, our participants reported that their reactions to this information being tampered with would differ accordingly. By using standardized abstract tasks representing the attack patterns, the current model lacks fidelity to the actual task environment and fails to accurately reflect the variability of human behavior adapting to the cybersecurity scenarios.

Undoubtedly, that task environment fidelity directly influences the model's performance in capturing human adaptability and dynamic interaction with the task environment. In addition, the fidelity of the task environment also constrains the precise simulation of perceptual-motor activities. As Fitts' law is embedded in the ACT-R architecture, the model is sensitive to the spatial properties of the task environment, which impacts the model's precision in simulating the processing time of perceptual-motor behavior (e.g., visual search and motor localization), especially in time-critical tasks. Task environment fidelity can also impact the model's effective application scope, as rule construction and necessary declarative knowledge must fit well with the setup and resources available in the task environment. Therefore, an effective cognitive model must sufficiently reflect the fidelity of the task environment to enhance the model's applicability and transferability to real-world scenarios.

### **3.3.3 Defining the Effective Task Environment of the Modeling's Application**

The third challenge is to define the effective task environment of the model's application.

In our exploratory modeling, participants mentioned their potential responses to cyberattacks beyond in-vehicle operations. Given the cognitive demands of processing attack-related information while maintaining driving safety, participants noted they might ignore a threat if it did not present as an immediate critical emergency. If an anomaly was perceived as unexpected or difficult to assess within drivers' capabilities or the timeframe, some participants indicated they would choose to pull over. In other words, instead of continuing to infer and respond to the threat while driving, drivers might choose to disengage entirely from vehicle operation, taking actions beyond the modeled scope of driving task (e.g., exiting the vehicle). The difference between these participants' decisions and our modeling assumptions that constrain driver responses within the driving operations raises questions about the effective application scope of such an in-vehicle driving task cognitive model for handling CAV cybersecurity.

The other side of the model's application scope is that cyberattacks differ from conventional vehicle system malfunctions, as they may be stealthy and less immediately perceptible to human drivers (Linkov et al., 2019; Nikitas et al., 2022; Rudd et al., 2017). Intentional deception is a common feature of cyber threats (Nikitas et al., 2022), making them harder to detect. Modeling how humans respond to stealthy and disguised attacks, different from the more salient and easily perceived ones, poses a particular challenge, as ACT-R-based models are more effective at predicting human reaction to clear perceptual stimuli than at capturing the more nuanced diagnosing process under uncertainty or unawareness.

Overall, identifying the appropriate modeling scope is crucial for validating the model's effectiveness and maximizing its capabilities. Therefore, the third challenge of applying an effective cognitive model in cybersecurity lies in identifying the effective task environment of the model's application, which requires balancing the analysis of domain-specific constraints with an understanding of the model's applicability.

## **3.4 Future Work**

Future research could also add more precise specifications of the in-vehicle environment and clear descriptions of attacks in perceptual terms, such as how altered cues, visual anomalies, or misleading

interface elements are presented and interpreted by drivers. Such future work could also enable investigation of how cybersecurity-related stressors push human cognitive capacity to its limits, offering a pathway to capture high-stress performance effects that the ACT-R model cannot directly simulate.

### **3.5 Summary of Chapter 3**

In this chapter, we introduced the leading classic cognitive architectures and selected the ACT-R model for application in CAV cybersecurity due to its widespread use, strong validation, hybrid structure, and comprehensive functional cognitive modules. Its adaptability and dominance in driving performance modeling further support this selection. We then tried applying one of the most used ACT-R models to simulate distraction-based cyberattacks targeting human drivers. Although the modelling effort was informal, it revealed several challenges in using the model to predict human vulnerabilities in responding to cyber threats.

First, the construction and mapping of production rules are often guided by the modeler's experiences and the task-specific analysis. While this approach may be applicable for modeling skill-based perceptual-motor tasks, it falls short when representing human information processing and analytic decision-making pathways that involve greater domain-specific knowledge and systematic analysis of the complex work domain. Thus, in cybersecurity, the simple and straightforward rule construction approach may fail to reasonably and reliably simulate human performance.

The second challenge of using the ACT-R model lies in its reliance on abstract psychological tasks representing the task environment in a simplified manner. While these subtasks induce perceptual and cognitive responses, they do not adequately reflect the complex, information-rich nature of cybersecurity task environments, which can fundamentally reshape human decision-making and performance. Inadequate representation of the task environment compromises the model's feasibility and reduces the precision of its simulation outcomes.

The third challenge is determining the effective task environment of the model's application. Humans' responses to various cyberattacks, whether distraction-based or stealthy, whether occurring within the vehicle or beyond, or no response at all, remain uncertain. This makes it particularly challenging to apply a cognitive model to a universal scope. Every cognitive model has its own strengths, areas for improvement, and inherent limitations for different practical applications. To fully

harness an ACT-R based model's strengths in simulating human performance in cybersecurity, an identification of the effective task environment is needed for the model's practical application.

Collectively, to answer our first research question:

- *RQ1: Is an ACT-R model sufficient for modeling human performance in cybersecurity?*

Several challenges must first be addressed before the ACT-R model can be fully applied to the cybersecurity domain. To tackle these challenges, the next chapter will explore how the Cognitive Work Analysis (CWA) can be integrated into the ACT-R based model and potentially address these challenges.

## Chapter 4 Towards Integrating CWA with the ACT-R Model

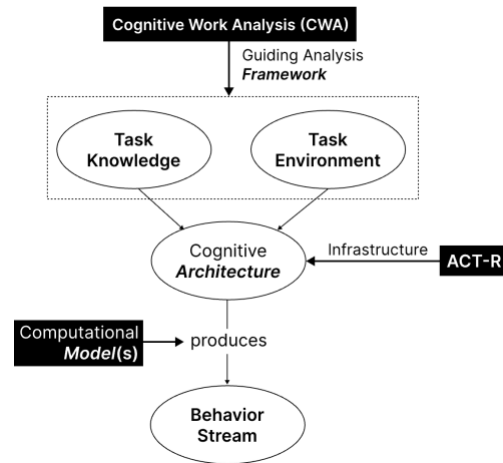
In this chapter, we introduce the Cognitive Work Analysis (CWA) framework and its strengths in cognitive modeling. Then, we will explain how each dimension of CWA can be integrated with the ACT-R model construction, potentially addressing the identified challenges in applying it to modeling human performance in complex domain such as cybersecurity, and answer the second research question:

- **Research Question 2:** *Can insights from Cognitive Work Analysis (CWA) enhance the effectiveness of an ACT-R cognitive model in complex domains such as cybersecurity?*

### 4.1 Cognitive Architecture, Model, and Analysis Framework

Before further exploring the integration of CWA and ACT-R, it is essential to clarify their fundamental differences: CWA is a cognitive analysis framework (Vicente, 1999), while ACT-R is a cognitive architecture (Anderson et al., 2004). As a framework, CWA provides a valuable structure for gathering domain-specific resources, but it does not necessarily lead to the development of direct answer addressing specific questions (C. Burns, 2013; Fidel & Pejtersen, 2004). In contrast, ACT-R offers a general infrastructure for simulating human cognitive processes through computational models. An illustration of how CWA and ACT-R as modelling approaches function differently for human performance modeling is shown in Figure 10.

Both CWA and ACT-R will be generally referred to as '*cognitive modeling approach*' in the following of this thesis; the term emphasizes their applications in informing the construction of a computational cognitive model, but not in serving directly as one.



**Figure 10.** Cognitive Architecture, Model, and Framework and How They Function within Human Performance Modeling (The original illustration is from (Byrne, 2012)).

## 4.2 Introduction to Cognitive Work Analysis (CWA)

CWA (Rasmussen & Jensen, 1974; Vicente, 1999) was born out of the control needs of the nuclear energy field, and is powerful in modelling complex social-technical domains so that flexible information systems that support human cognition can be developed (Wurst et al., 2021). It is widely applied across many complex socio-technical domains, such as in healthcare, nuclear power plants, transportation systems, and cybersecurity of financial sector organizations (Burns et al., 2008; Hulme et al., 2019; Knight et al., 2018; Kovesdi & Spielman, 2021; Schmid et al., 2020; Van Der Kleij et al., 2022). CWA is particularly good at analyzing the specific work environment by considering the constraints that affect a system's capacity to fulfill its purpose (Birrell et al., 2012).

Numerous applications of CWA in cognitive research have demonstrated its strengths in structuring domain-specific analyses to support and enhance human decision-making. The work conducted by Birrell et al. (2012) started with aligning information presentation with human cognitive processes when designing interfaces for vehicles. Stanton and Allison (2020) extended this work by applying the complete CWA framework to inform an environmentally friendly driving assistance system design, emphasizing information provision to enhance fuel efficiency. Seppelt and Lee (2007) used CWA-derived Ecological Interface Design (EID) (C. M. Burns & Hajdukiewicz, 2017) to represent Adaptive Cruise Control (ACC) decision-makings visually, guiding appropriate interpretations of ACC and facilitating smooth transitions between manual and automated control. An

interesting study by Cornelissen et al. (2013) investigated the variability in human decision-making at road intersections. The authors applied WDA, ConTA, and Strategy Analysis to understand the real-world contextual interactions among drivers, motorcyclists, cyclists, and pedestrians in negotiating stops or entering the intersections with different groups of road users. This work clearly showcases CWA offers a systematic perspective on multifaceted road interactions and the system's dynamic complexity: different road user groups share and compete for limited resources (e.g., road space and infrastructure) to fulfill their respective goals (e.g., safety and space). In a recent CWA research (Zhang & Lintern, 2024), the authors mapped different traffic scenarios into CWA for SAE automation Level 0 to 4 AVs. The effort aimed to identify the conflicts among driving values (e.g., safety, productivity, multi-tasking, navigation, and efficiency) in the current ADS system design. This work highlights again the key strength of CWA in providing a systemic view of a specific domain, rather than treating subsystems as isolated components without considering their interactions and trade-offs in real-world functioning.

In summary, CWA's core strengths lie in its classic top-down framework (see Figure 11) in guiding both interactive (e.g., considering different road users' perspectives) and systematic analyses (e.g., tracing from overarching system goals to specific functions such as fuel efficiency) of highly complex socio-technical systems. This makes it a compelling approach for analyzing human decision-making within multifaceted, dynamic environments constrained by domain-specific factors.

### **4.3 Towards an Enhanced Model through the Integration of CWA**

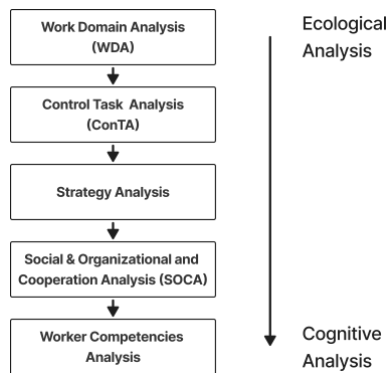
As noted in Chapter 3, the primary challenges in constructing the ACT-R model and applying it to complex domains, such as cybersecurity, include the lack of task-specific knowledge and insufficient analysis of the task environment. CWA appears to be a highly complementary modeling framework for addressing these gaps. Drawing on its broad applications and proven success in research across diverse systems, CWA offers a structured approach to analyzing human reasoning within complex work environments from a domain-specific view. In turn, the ACT-R model provides quantifiable means to predict and validate the CWA insights integrated into the model's construction. By integrating both modelling approaches, the resulting model is expected to be an enhanced simulation tool, capable of capturing nuanced human reasoning and precise actions within complex, domain-specific environments, such as in cybersecurity.

To be more specific, the expected enhancement by the integration is: 1) To improve the ACT-R model's construction and adaptability to complex real environments in CAV cybersecurity with CWA's well-structured knowledge and work environment analysis framework; and 2) To inherit the computational power from ACT-R model as quantifiable assessment of humans' vulnerabilities in CAV cybersecurity.

Therefore, in the following, we will delve further into the compatibility of CWA and ACT-R for integration as an enhanced cognitive model.

### 4.3.1 Conceptual Overview of the Integration

This section provides a detailed outline of the conceptual integration of the two modelling approaches. CWA has a clear dimension-based framework, so this conceptual integration exploration here will be explained based on CWA's structure (see Figure 11). (Note: This chapter presents a conceptual integration proposal only. The detailed illustration of the integrated model construction and outcome validation will be provided in Chapter 6.)



**Figure 11.** Five Phases of CWA (Rauffet et al., 2015).

### 4.3.2 Abstraction Hierarchy (AH) and Declarative Knowledge

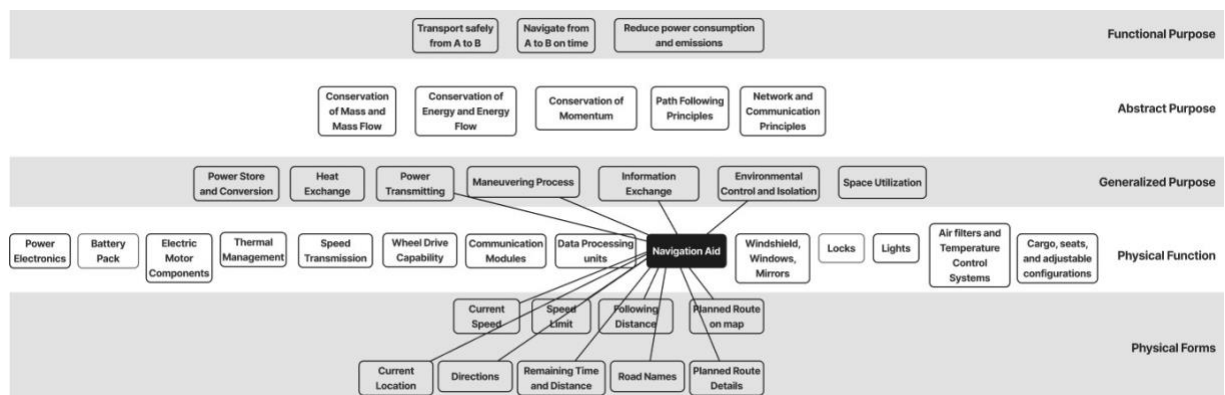
Work Domain Analysis (WDA) is the first dimension of CWA. It provides an event and actor-independent description of system properties. WDA is to identify a targeted scope of the work domain, which can be regarded as the “external world” in ACT-R's basic structure (see Figure 4).

The most used tool of WDA is the Abstraction Hierarchy (AH). AH uses means-ends relationships to describe the work domain through different levels (see Figure 12). The elements at

one level are used to achieve the elements at a higher level (Vicente, 1999), presenting a dynamic view shift of the system with the changing demands of tasks (Naikar, 2017).

An AH of a vehicle system (see Figure 12) was developed based on prior WDA research's AHs of the vehicle and transportation domain (Salmon et al., 2007, 2015, 2019; Zhang & Lintern, 2024), and further refined using the author’s knowledge of vehicle systems. This AH serves only as an illustrative example to support the conceptual integration discussed in this work. (This resulting AH was serving only as an example to support the conceptual integration’s explanatory purposes.)

In the two lower levels (physical forms and physical) of AH, there are more tangible physical elements that humans can operate on. These tangible physical elements thus compose the task environment for direct human interaction, while the means–end links can be reflected as task knowledge upon the environment. For example, the “Navigation Aid” (see Figure 12), as in the “Physical Function” Level, is connected to more tangible information in “Physical Forms”. This tangible information can thus be used to establish the task knowledge elements in the model.



**Figure 12.** The Abstraction Hierarchy of a vehicle system (Only a portion of the Physical Forms and connections is presented, as it is intended solely to serve as an AH example for explanations).

In ACT-R, declarative task knowledge is represented in structures called *chunks* (Bothell, 2022). Chunks are lists of attribute-value pairs, and each slot is a named attribute (*The ACT-R Cognitive Architecture and Its Pyactr Implementation*, 2020). The "Navigation Aid" in the AH, as Physical Functions with their connected Physical Forms, turns out to be structurally perfect for representation as a declarative chunk with attribute slots for acquiring dynamic slot values from the real task environment, as presented in Figure 13.

```
(chunk1
  isa ADAS
  name navigation_aid
  current_speed 55
  speed_limit 60
  following_distance 30
  remaining_time_distance 120
  road_name "Highway 401"
)
```

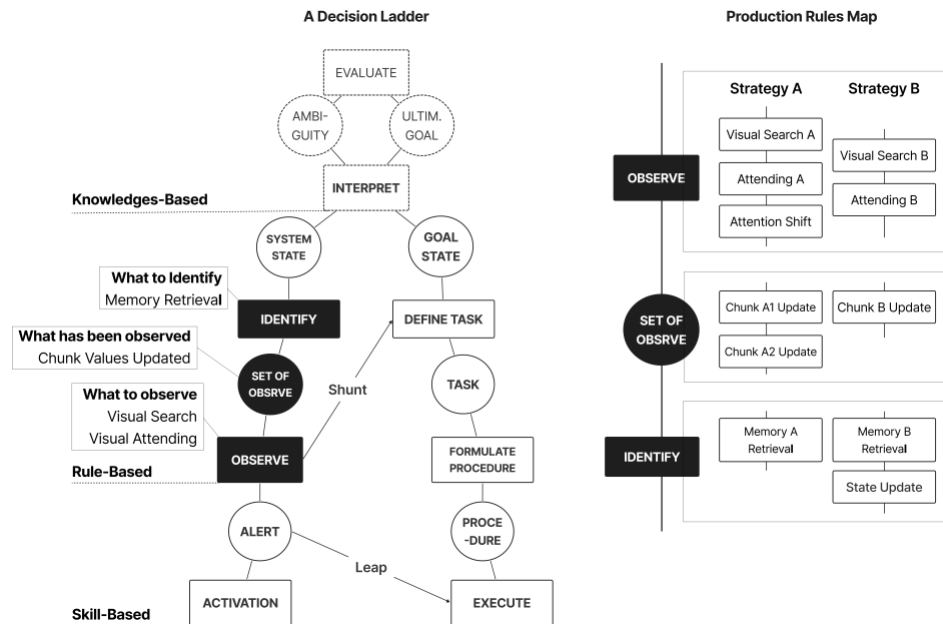
**Figure 13.** Declarative Memory Chunk Example (from an Abstraction Hierarchy of a vehicle system).

In this way, a WDA with AH can serve as a comprehensive domain-specific library for collecting sufficient task knowledge to enhance the model's task knowledge during declarative chunk preparation.

### 4.3.3 From Decision Ladder (DL) to Production Rules

Control Task Analysis (ConTA), CWA's second dimension, defines tasks as what needs to be done in the work domain (Vicente, 1999). Its tool, Decision Ladder (DL), is constructed by connecting two types of nodes: *boxes*, which represent *information processing activities*, and *circles*, which represent *states of knowledge*. A DL connecting these boxes and circles offers a non-linear template that generalizes the procedure to achieve the task goal.

In comparison, the ACT-R architecture consists of modules connected by production rules. Each module is dedicated to processing a distinct type of information. Perceptual modules are used for identifying perceived objects (e.g., an object in the visual field), a manual module is used for action control, and the goal module is used for tracking current goals and intentions (Anderson et al., 2004). The production rules define how changes in module states result in chunk value updates or motor actions. Through this process, the model could transform declarative knowledge stored into a procedural form (e.g., manual action or module content updates). Notably, this transformation process has similarities to the information processing activities and states of knowledge transformation as represented in the DL.



**Figure 14.** A Decision Ladder (DL) (Left) (part of the rule-based process shaded in black as an example to guide analytic rules connections); and Part of A Production Rules Map (Right) following DL’s template for rule set connections.

Notably, there is a difference in granularity between DL and production rules. Arguably, each transformation between boxes and circles in DL does not strictly correspond to a production rule. Production rules reflect a more decomposed transformation process, describing the process of updating chunk value updates or motor action triggering (see Figure 14 (Right)), which are finer than the DL’s identification of only input/output as states of knowledge and activities without the detailed process. A DL’s transformation thus could be further decomposed into a set of connected or parallel production rules (see Figure 14 (Left)). The distinction arises from the nature between the two: ConTA provides a top-down analytical template focused on what the task environment offers to achieve the goal, whereas production rules represent a bottom-up model detailing how humans allocate cognitive resources across different brain functional areas to interact with the task environment to approach the goal.

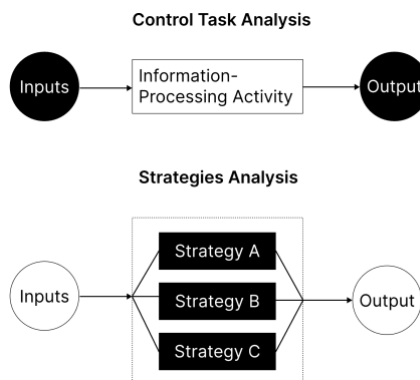
In addition, the DL’s segmentation based on the S-R-K taxonomy can inform the construction of production rules and the switching mechanisms under conditions of competing cognitive resources. Specifically, in the DL, the skill-based segmentation involves more perceptual-motor action rules, while the rule-based part engages more nuanced inferences and analytic processes. Compared to skill-

based decisions, which rely primarily on direct perceptual-motor mechanisms, rule-based processes (see Figure 14) require more specific task knowledge to construct rules supporting analytic processing. Given that, the model's analytic production rules can follow the DL's template as procedural guidance for constructing analytic-based rules, rather than informal heuristic rule construction approaches without procedural guidance. Even shortcut paths within the DL, such as *shunts* or *leaps* (see Figure 14), may guide adaptive learning mechanisms inherent to ACT-R's production compilation process. In this sense, ConTA contributes primarily to coordinating the connections of production rules along different pathways toward task completion following a systematic template.

Collectively, the *Decision Ladder* provides overall guidance on how production rules can be structurally coordinated and mapped to achieve ultimate task goals, especially within complex task environment with intensive analytic process, but does not determine precisely how individual production rules should be constructed.

#### 4.3.4 Strategy Analysis and Production Rules Selection

When using the DL to guide the coordination and connections of production rules with input and output states, the process by which the inputs are transformed into outputs is not explicitly detailed. In other words, the DL alone does not specify how particular conditions (i.e., IF-) trigger specific rule firings (i.e., THEN-). The third dimension of CWA, known as Strategy Analysis, addresses this gap by specifying our pathways for transitioning between input and output nodes within a DL (see Figure 15), which could underlie the establishment of specific condition-action pairs for production rules.



**Figure 15.** Comparison of ConTA and Strategies Analysis (Vicente, 1999).

The great flexibility of Strategy Analysis insights could also contribute to the rules-switching mechanism. A strategy, as defined by (Rasmussen, 1983), is a category of cognitive task procedures that transform an initial knowledge state into a final knowledge state. Instead of a specific instance of a procedure, a strategy analysis is a category of procedures. This concept offers flexibility to accommodate human adaptability and dynamic learning into the model. ACT-R's production rules selection and switch (Bothell, 2022) by the utility reward mechanism could also draw upon insights from Strategic Analysis's pool of candidate rules and switching rationale accordingly.

#### **4.3.5 Overview of the Integration**

CWA and ACT-R differ in their analysis scopes, emphasis (CWA focuses on work environments, ACT-R on human cognition capability), validation approach (CWA emphasizes domain insights, ACT-R relies on empirical objective data), and outcome forms (ACT-R is more quantitative, CWA is more qualitative).

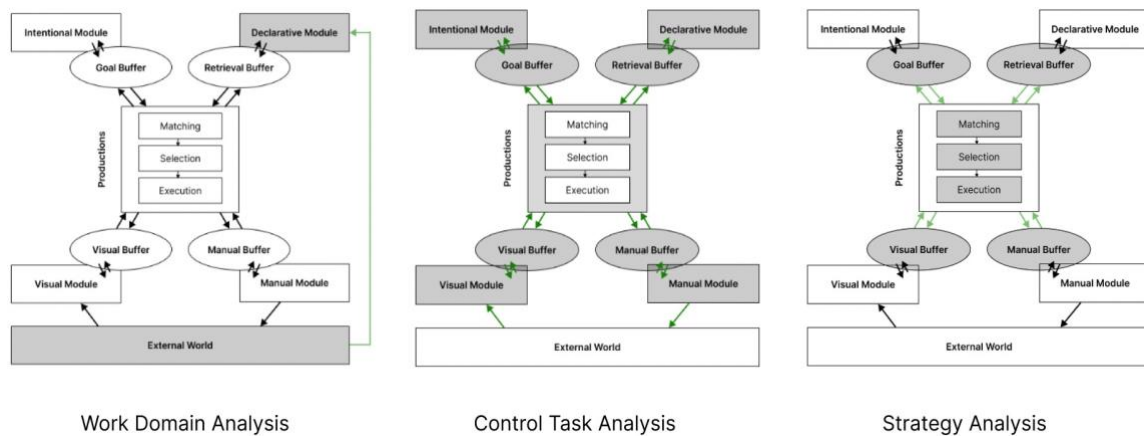
However, CWA and ACT-R, as cognitive modeling approaches, still share significant similarities in their fundamental structure and concept of cognition (see Figure 16). CWA's key dimensions, such as WDA, ConTA, and Strategy Analysis, can effectively guide the collection of detailed information about the work environment's constraints and domain-specific knowledge for human decision-making. These insights are a valuable resource for enhancing the construction of an ACT-R-based model with sufficient task knowledge and a high-fidelity representation of the task environment. In turn, ACT-R's modeling outcomes, including the fine-grained temporal sequence of behavioral events and quantitative measures of human behavior, can provide quantifiable evidence for CWA-based insights.

Together, the integration outlined above presents a promising, enhanced computational modeling approach to address the identified challenges in simulating human performance within complex systems, such as in cybersecurity applications (see Section 3.3).

#### **4.4 Contribution**

To the best of our knowledge, this study represents the first attempt to examine the compatibility of ACT-R and CWA for developing an integrated cognitive modeling. As a conceptual analysis, this chapter aims to provide a promising direction for systematically translating CWA-guided qualitative

insights into computational simulations of human performance grounded in the ACT-R architecture. Although this chapter presents only a conceptual analysis, the proposed integration shows potential to strengthen the modeling of human performance in cybersecurity and other complex work domains.



**Figure 16.** CWA's Key Dimensions within ACT-R Architecture.

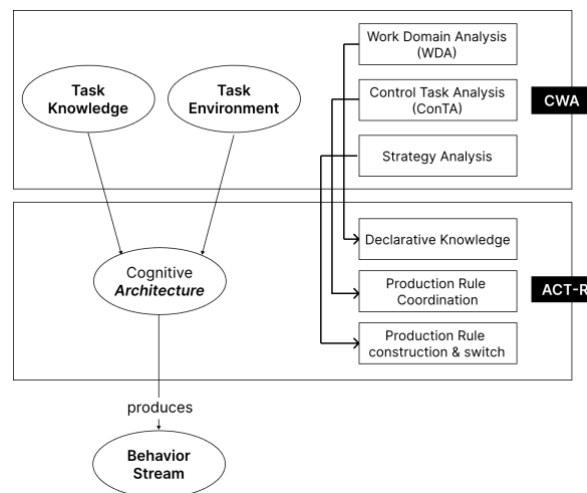
## 4.5 Summary

Here, we combine the cognitive modeling challenges identified in Chapter 3 with Chapter 4's conceptual exploration of integrating CWA insights into ACT-R model construction to answer the second research question:

- *RQ2: Can insights from Cognitive Work Analysis (CWA) enhance the effectiveness of an ACT-R cognitive model in complex domains such as cybersecurity?*

Chapter 3 revealed three primary challenges when directly applying an ACT-R-based model to a complex work domain: a lack of domain-specific task knowledge analysis, low fidelity representation of the task environment, and the need to identify a suitable task environment. These insufficiencies thus question the direct application of an ACT-R model in complex, information-intensive, and highly dynamic task environments, such as in cybersecurity. To address these challenges, we introduce CWA as a robust cognitive analysis framework with a strength in investigating how domain knowledge and work environment constraints influence human decision-making and propose integrating it into the model as a complementary enhancement.

We then explored the compatibility of the three dimensions of CWA with the construction of the ACT-R model. Work Domain Analysis (WDA) provides domain-specific functional and reasoning-based knowledge, serving as a task declarative knowledge library. Control Task Analysis (ConTA) connects the domain knowledge to task-specific goal achievement for rules coordination. Strategy Analysis captures the underlying variability and flexibility in human decision-making, offering a pool of parallel rules and rule-switch strategies for applying the model's utility mechanisms in a more domain-specific and validated manner. Together, these CWA dimensions support a more structured analysis of task knowledge and a more dynamic representation of the realistic task environment. Hence, to answer the question, the integration of CWA with ACT-R is expected to enhance modeling effectiveness, accordingly, as illustrated in Figure 17.



**Figure 17.** The Cognitive Model by the Integration of CWA and ACT-R.

Next, before implementing the enhanced ACT-R model with CWA insights, we need to resolve the challenges in identifying the effective task environment for cybersecurity management to apply the model. Accordingly, the next chapter will examine whether current in-vehicle human operation is an effective task environment for detecting and mitigating cybersecurity threats, and what other task environment within the CAV cybersecurity defense framework would be more effective in managing cybersecurity threats to apply the proposed cognitive model.

## Chapter 5 Cognitive Model Application: Assessing In-Vehicle Human Responses to CAV Cyberattacks

The third cognitive modeling challenge, as discussed in Chapter 3, is identifying an effective task environment (Section 3.3.3). We will use the in-vehicle operation as a continuing example to analyze and identify how the task environment impacts drivers' analytic decision-making and effective behaviors in responding to cyberattacks and examine whether the enhancements of the CWA-informed ACT-R-based model in predicting human performance could be appropriately applied in this use case. Specifically, this chapter examines how humans in-vehicle process CAV cyber threats, leading to their ultimate responses, and whether the in-vehicle humans' responses are effective in detecting and mitigating cyber threats, answering the research question:

**Research Question 2a:** *Is in-vehicle operation an effective task environment for detecting and mitigating cybersecurity threats?*

In study 1, we will answer this question by first examining human drivers' detection of different forms of cyberattacks and how the in-vehicle user interface displays information may influence their awareness. A survey study will be conducted to capture the broader attitudes and experiences of drivers. The survey study's anticipations are:

- 1) Analyzing drivers' awareness in different cyberattacks.
- 2) Evaluating the effectiveness of the current in-vehicle interface in supporting drivers' awareness and effective response to cyber threats and providing improvement suggestions.

In addition, these findings will be used to assess the effectiveness of in-vehicle responses to cyberattacks, inform the discussion on effective task environments in managing cyberattack threats, and for the model's application.

### 5.1 Study 1: Driver Responses to Silent and Explicit Cyberattacks with In-Vehicle Display Interfaces

#### 5.1.1 Introduction

The increasing complexity of driving systems has introduced numerous vulnerabilities. Many safety-critical incidents can be traced back to cybersecurity concerns (Jongen et al., 2016; Malik & Sun,

2020; Payre et al., 2023) and the potential exploitation of weaknesses within Automated Driving Systems and ADAS (Alsaade & Al-Adhaileh, 2023; Bachute & Subhedar, 2021; S. Kim & Shrestha, 2020; Xie et al., 2021). In particular, the causes of cyber threats are not solely rooted in technological vulnerabilities but also from evolving driver interactions with ADAS, where lower engagement and distractions (Ban & Jeon, 2025; He & Burns, 2022), and limited transparency (Luo et al., 2025) introduce additional risks. Some studies focus on cyber threats targeting the user-centric interface of vehicle systems (Linkov et al., 2019). The sheer volume of data, along with the specialized knowledge and experience required, makes it difficult for general in-vehicle human operators to respond effectively and promptly to cybersecurity threats. These challenges are further compounded by the inherent vulnerabilities of highly interconnected CAV systems, where information is often opaque, attacks may take deceptive forms, and drivers have limited knowledge and restricted access to the broader network. As a result, managing cybersecurity threats from within the vehicle becomes significantly more difficult for human operators.

From Chapter 3's example, we observed that human drivers may exhibit different decision-making patterns in response to various types of cyberattacks. Specifically, we strive to investigate how drivers process in-vehicle information presented through the console display and how they detect different forms of cyberattacks, which are manifested through anomalies in vehicle motion and displayed information.

## **5.1.2 Background**

### **5.1.2.1 In-Vehicle Cybersecurity Situation Awareness (CSA)**

CSA is a popular topic in human factors research on cybersecurity, which applies the situation awareness (SA) concept to the cyber domain, examining how well humans or systems detect, understand, and anticipate cyber events (Jaeger & Eckhardt, 2021; Sawyer & Hancock, 2018). Unlike the more established scales for measuring SA, CSA assessments require a close alignment with domain-specific contexts and objectives to ensure accuracy and relevance. Research on CSA in the context of cyber threats to driving systems remains challenging due to the complexity of the driving environment and the diversity of ADAS functionalities (Zhou et al., 2022). Moreover, measuring CSA is particularly difficult because cyberattacks could be intentionally deceptive and remain stealthy, making them harder to detect (Nikitas et al., 2022; Rudd et al., 2017). This issue is one of the key characteristics that distinguishes cyberattacks from conventional vehicle malfunctions. Therefore,

evaluating human drivers' awareness will also consider the possibility of not being able to perceive the stealthy anomalies firsthand.

Payre et al. (2023) thus introduced a new perspective by examining two distinct categories of cyber-attacks for human drivers' awareness: Explicit (e.g., a ransomware attack displayed on the in-vehicle screen) versus Silent (e.g., inactive turn signals without any alert on the vehicle's screen or instrument cluster). Their study used a driving simulator with 38 participants to examine two types of attack: an explicitly displayed ransomware message and a silent turn-signal malfunction. The results show that nearly half the participants spent over 12 seconds looking at the ransom message, while most failed to detect the silent failure. This finding further reinforces the concern that general drivers may not detect silent and stealthy cyberattacks within the in-vehicle environment. Building on this, Study 1 is motivated to explore the varying levels of attack explicitness and how drivers perceive, interpret, and respond to cyberattacks with different levels of salience. Payre et al. (2023)'s definitions of the explicit and silent attack categories were retained in our experimental design.

#### 5.1.2.2 In-Vehicle Information Display in Cybersecurity

In-vehicle displays and infotainment systems are integral to the driving experiences (Dudziak et al., 2024). These functional features enhance the driving experience by offering rich entertainment services and expanded displays that provide drivers with more information. Executed well, these systems enhance safety by improving the SA of the driver. However, the in-vehicle displays and services have heightened overall vulnerability, as these displays and services are also possible targets for cybersecurity attacks (Martínez-Cruz et al., 2021).

Moreover, the interaction between drivers and the in-vehicle interface, which displays information in the event of cyberattacks, requires further analysis. Findings from Study 1's results suggest that human cognitive resources can be primarily occupied by the analysis process of driving-related information, increasing vulnerability to emergency takeover requests. However, how drivers process and interpret different types of information, including driving conditions, vehicle status, and potential threat cues, during the detection and mitigation of cyber threats has not been deeply investigated. Accordingly, this study will further examine how drivers utilize a representative in-vehicle user interface to capture information and detect cyber threats, focusing on the characteristics of the displayed information that influence human drivers' identification, interpretation, and decision-making of cyber-attacks.

Tesla is undoubtedly a prominent name in automated driving (Backlinko Team, 2025) and has led to a new era and trend for ADAS interface design (Song, 2021). The Tesla Model 3's redesign of the dashboard shift towards minimal physical controls and an enlarged central touchscreen, has been praised for its enhanced road visibility and intuitive presentation of information (Gillmore & Tenhundfeld, 2020). At the same time, this redesign raises concerns about large-screen distractions and poor transparency of system communication (Gillmore & Tenhundfeld, 2020; Song, 2021). Given the mixed feedback on its design, the Tesla Model 3's console display was selected as the experimental display in our study to understand the efficacy of the console display for cyber threat identification, not as a critique of Tesla's console design but as a widely used example of console displays in contemporary vehicles, and, in many ways, a best-in-class example.

### 5.1.3 Hypotheses

Our hypotheses for Study 1, which are based Payre et al. (2023)'s work, are summarized below.

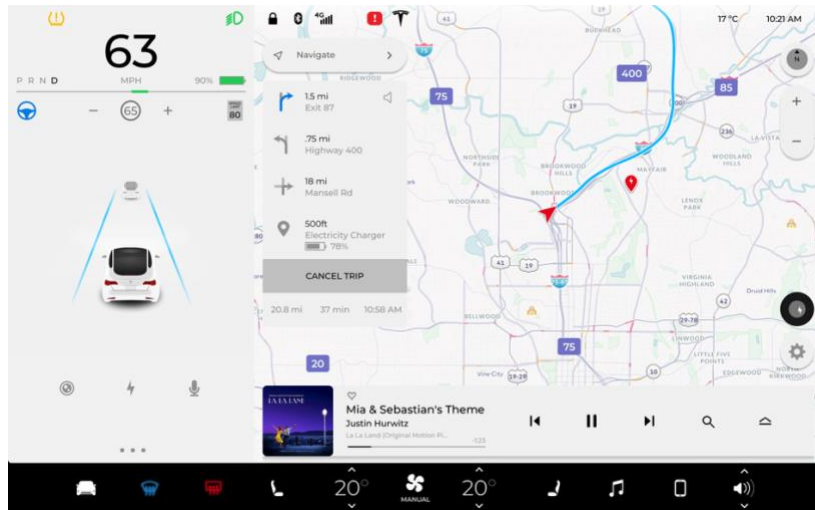
- **H1:** Drivers will exhibit greater awareness of more explicit cyberattacks than silent ones.
- **H2:** The limited usefulness of current in-vehicle displays hinders drivers' ability to gain a full awareness of cyber threats.
- **H3:** Inadequate awareness of cyberattacks can lead to inappropriate responses and heightened risks to driving safety.

### 5.1.4 Experiment Design

#### 5.1.4.1 Explicit vs. Silent Attributes

Study 1's experimental design primarily emphasized the differentiation of explicit and silent attributes (Payre et al., 2023) by narrowing to three distinct scenarios, each marked by noticeable changes in elements on in-vehicle HMIs and motion status within the driving context. Since this study does not aim to develop new interface solutions or evaluate deficiencies in current display designs, but rather to examine how human drivers respond to different forms of cyberattacks presented through the prevalent in-vehicle display, no modifications to the interface (Tesla Model 3's) were made or proposed. This study uses Tesla Model 3's original alert sounds, and interface elements (e.g., navigation, collision-avoidance alerts, map objects, and route indicators) implemented through the

Figma asset set of Tesla ([Link](#)) (see Figure 18). The vehicle motion status followed the same road conditions used in Chapter 3 and retained the first-person driver view during lane-change events.



**Figure 18.** Example preview of the Tesla Model 3 interface prototype employed in this study, created using Figma design assets ([Link](#)).

All three attack scenarios targeted a Level 2 automated vehicle, with participants experiencing the simulated attacks from the perspective of a first-person driver. Each scenario is designed to replicate prevalent cyber threats (Khan et al., 2020) targeting critical ADAS functions, including navigation and collision avoidance. The scenarios' details are described as follows:

- **Scenario 1:** The vehicle deviated from its intended path without visible alarms or cues on the console display or driving environment, making this a Silent attack. This anomaly was primarily caused by GPS spoofing (Abrar et al., 2024; Tzoannos et al., 2024), where false signals mislead the vehicle's GPS receiver, leading to a risky route and direction. In this scenario, the navigation system displays misinformation on the route when the vehicle is intended to continue along a straight highway. The inconsistency arises because the vehicle motion status maintains a straight trajectory, yet the navigation system indicates a right turn. This silent scenario thus represents the displayed route information contradicting the actual driving conditions without explicit cues.
- **Scenario 2:** The collision avoidance system activated upon detecting a close leading vehicle with auditory (Tesla Standard collision avoidance warning sound) and visual (leading vehicle

highlighted in red on the display’s map and views) alarms. The alarms turn off after a few seconds but fail to alert to another approaching vehicle soon after. The status changes of the auditory and visual alerts of the collision avoidance system make this scenario Explicit. This attack may result from unauthorized ECU firmware modifications (Lopez et al., 2019), and the resulting anomaly poses a potential risk of undetected vehicle collisions.

- **Scenario 3:** Similar to Scenario 2, the collision avoidance system detects a leading vehicle and triggers alarms. After a brief deactivation, the display erroneously signals a non-existent vehicle, potentially causing the car to steer abruptly. The first-person driving view could also observe the leading vehicle on the roadway, and its abrupt lane change. The motion status change, continuous auditory and visual alarms make this scenario the most explicit condition.

To better illustrate the differences among the scenarios, Figure 19 presents screenshots of the three designed attack scenarios, and their levels of explicitness are evaluated in Table 3.

**Table 3.** The "Explicitness" of the Attacks in Scenarios.

Scenarios	1	2	3
<b>In-Vehicle Information Cluster Display(s)</b>			
New Object(s) Appearance	N	Y	Y
Auditory Alert(s)	N	Y	Y
<b>Vehicle Motion Status</b>			
Lane Changing	N	N	Y
<b>Overall Explicitness</b>	<b>Low</b>	<b>Medium</b>	<b>High</b>

## 5.1.5 Methods

### 5.1.5.1 Participants

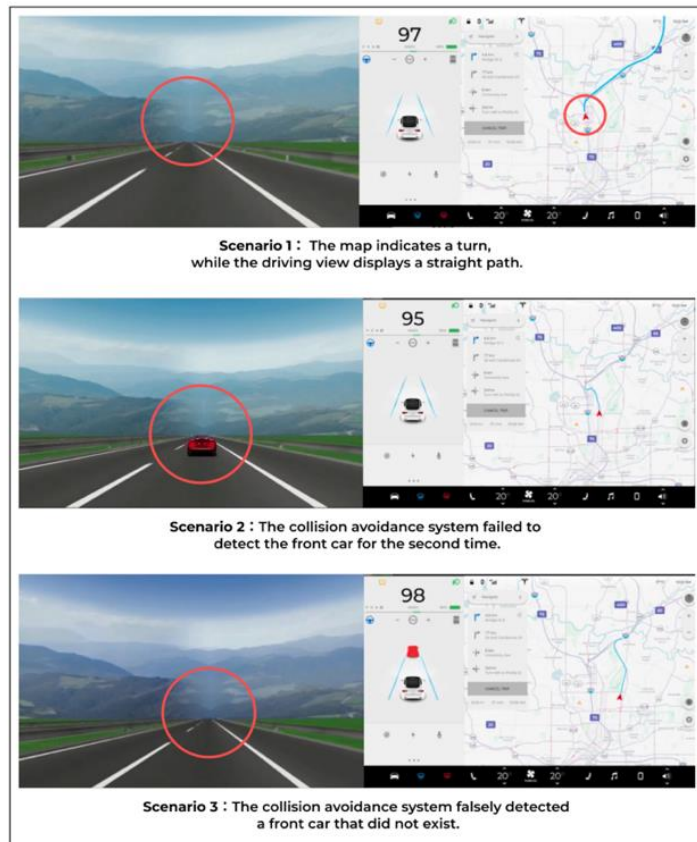
Ninety-four adult participants were recruited via Amazon Mechanical Turk for our online study conducted on Qualtrics®. Given the focus on driving tasks, possessing a valid driving license was a prerequisite for inclusion. Additionally, the study’s emphasis on detailed information observation and decision-making necessitated a minimum screen size of 13” for participation. Participants were compensated with a \$10.00 stipend through AMT’s payment system.

To collect a large, diverse sample of human drivers' responses across regions, cultures, vehicles, and driving habits, the study was designed as an online survey on a crowdsourcing platform

(Paolacci & Chandler, 2014). However, recent work has shown that such platforms are increasingly affected by unqualified workers and automated or bot-generated responses (Kennedy et al., 2020). To mitigate these concerns of response qualifications, we restricted recruitment to Master workers with the overall HIT approval rate greater than 97% on AMTurk as the eligibility criteria.

To further reduce the likelihood of automated responses on the crowdsourcing platform and to ensure data quality, an attention-check item was embedded and disguised as one of the attack scenarios. Midway through the video, a large-font message appeared stating, “Attention Check: Please answer ‘No’ to the first question below.” Participants who did not respond “No” to that question were classified as providing invalid responses and were excluded from subsequent analyses.

This research received ethical approval from the University of Waterloo’s Ethics Committee.



**Figure 19.** Screenshots of Experiment Videos of Cyber Attacks.

### 5.1.5.2 Procedure and Measurements

Participants began the study by completing a demographic questionnaire that captured details such as age, driving experience, driving habits, and familiarity with ADAS usage. They then watched an introductory video of the Tesla Model 3 console to familiarize those unfamiliar with it, followed by four online scenarios that included three experimental videos and one attention check, which helped identify cybersecurity threats.

Each attack scenario video lasted approximately 30 seconds and was embedded with a prototype interface (Johnson, 2017) of the Tesla Model 3's driving assistance console, displaying driving-related data and embedded attack cues in real-time. Immediately after each video, participants' responses were collected and evaluated across three dimensions:

***Anomaly Awareness and Interpretation:*** Participants first answered whether they had noticed any anomalies that could potentially indicate a cyber threat. If the participant answered "yes," a follow-up question asked them to infer the potential cause of the anomaly. A response was considered correct if the selected option included the designed anomaly. The third question assessed how the participant would potentially respond to the detected anomaly. After completing their responses, participants were shown the explanation of the designed anomaly and its cause to verify their assumptions.

***Display Usefulness Assessment:*** Following the first part, participants evaluated the interface's effectiveness in helping them become aware and identify the anomaly's potential origins using a 7-point Likert scale, where 1 represented "Not Useful" and 7 signified "Very Useful".

Finally, participants were asked an open-ended question about any additional information they believed was necessary for identifying and responding to the threat, and any feedback or concerns they had about the study.

### 5.1.6 Results Analysis

A within-subjects design was employed to investigate the impact of the three designed attack scenarios on participants' performance in anomaly detection and interpretation, their evaluation of the console display's usefulness, and a summary of their expected response and suggested improvements to the console display.

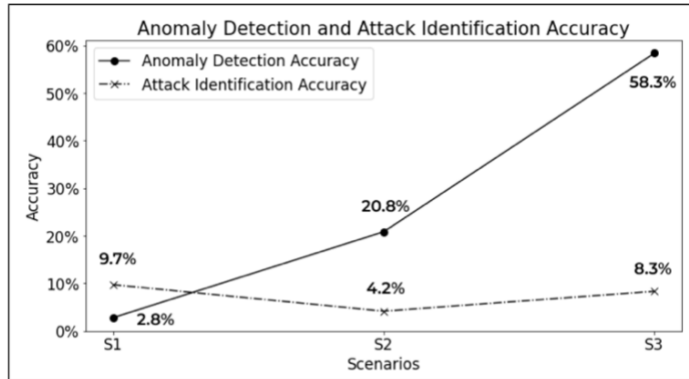
### 5.1.6.1 Demographic

Seventy-two qualified participants' responses (aged 18 to above 70) contributed to the results, with the demographic distribution as follows: 47.2% aged 30-39, 23.6% aged 40-49, 15.3% aged 50-59, and 11.1% over 60. Of these, sixty-five (90.3%) participants possessed a driving license for more than 10 years, with 63 engaging in driving at least weekly, indicating that most of our participants are experienced and active drivers. A majority (46 participants, 63.9%) acknowledged possessing concern about cybersecurity issues within driving contexts, yet self-identified as novices with minimal or no expertise in this domain.

### 5.1.6.2 Anomaly Detection and Attack Identification Accuracy

A Cochran's Q test was conducted to analyze the differences in participants' detection rate scenarios and interpretation correctness across the three scenarios. Participants' performance in anomaly detection accuracy demonstrated significant differences across three scenarios ( $Q(2) = 25.4, p < 0.01$ ). Post hoc McNemar's tests (Bonferroni-adjusted  $\alpha = 0.05/3 = 0.017$ ) revealed significant differences in detection accuracy between Scenario 2 and Scenario 3,  $\chi^2(1, N = 72) = 4.00, p = .007$ , and between Scenario 1 and Scenario 3,  $\chi^2(1, N = 72) = 1.00, p < .001$ . The comparison between Scenario 1 and Scenario 2 did not reach the corrected threshold for significance,  $\chi^2(1, N = 72) = 11.00, p = .029$ . The post-hoc tests indicate that Scenario 3 elicited significantly higher detection rate, as shown in Figure 20.

However, the level of correct reason identification did not exhibit significant differences ( $Q(2) = 1.73, p = 0.42$ ).



**Figure 20.** Anomaly Detection and Attack Identification Accuracy Across Three Scenarios. Participants' performance in anomaly detection accuracy (solid line) demonstrated significant differences across three scenarios ( $Q(2) = 25.4, p < 0.01$ ); however, the level of correct reason identification (dashed line) did not exhibit significant differences among three scenarios ( $Q(2) = 1.73, p = 0.42$ )

#### 5.1.6.3 Display Interface Usefulness Assessment

A Friedman test revealed a significant difference in participants' assessed Display Usefulness in detecting and interpreting cyber threats across scenarios,  $\chi^2(2, N = 54) = 6.70, p = .04$ , with a small effect size (Kendall's  $W = .062$ ). However, Bonferroni-corrected Wilcoxon tests revealed no significant pairwise differences ( $ps > .017$ ). Scenario 3 shows a relatively higher score for the usefulness of the display supporting anomaly detection and identification (Scenario 1 mean score = 3.33, Scenario 2 mean score = 3.37, Scenario 3 mean score = 3.89). This result indicates that, although perceived usefulness varied across scenarios overall, the display was not rated as significantly more helpful in any specific scenario compared to the others.

#### 5.1.6.4 Interface Improvements and Information Requirements

When asked about additional information on console display to improve the anomaly detection capabilities, most participants (30 in Scenario 1, 20 in Scenario 2, and 15 in Scenario 3) preferred alarms with more salient visual (e.g., pop-up error messages) and auditory cues.

Conversely, some participants (5 in Scenario 1, 11 in Scenario 2, and 12 in Scenario 3) deemed the existing display interface's information sufficient, as shown in Table 4, attributing detection failures to a lack of experience and attention rather than information inadequacies.

Additionally, a few respondents desired more comprehensive diagnostic data, including network and sensor status, to address potential threats better (1 in Scenario 1, 14 in Scenario 2, 16 in Scenario 3). This indicates some driver’s demand for deeper system component status information to improve their comprehension of cybersecurity threat detection. Participants also emphasized the importance of receiving guidance on subsequent actions as crucial information at the moment of detection (1 in Scenario 1, 4 in Scenario 2, 2 in Scenario 3).

**Table 4.** Console Display Interface Improvement Suggestions.

Scenarios	1	2	3
Alert Improvements	30	20	15
The displayed information is sufficient	5	11	12
Sensors & Network status, and Diagnostics Information	1	14	16
Guidance on subsequent actions	1	4	2

#### 5.1.6.5 Initial Response Actions

Across all scenarios, the top-rated first action response to perceived anomalies was to pull over safely and restart the system (25 in Scenario 1, 32 in Scenario 2, 48 in Scenario 3), followed by initiating onboard diagnostics (18 in Scenario 1, 27 in Scenario 2, 34 in Scenario 3), as shown in Table 5.

**Table 5.** Chosen Initial Action(s) If Encountering the Anomalies.

Scenarios	1	2	3
Safely pull over and restart the vehicle	25	32	48
Run any available diagnostic tests via the vehicle’s onboard computer	18	27	34
Report the issue to the vehicle manufacturer or authorized service center	17	13	34
Consult the vehicle’s user manual for troubleshooting	14	12	19
Immediately turn off the affected system if possible	6	15	31
Manually inspect the systems for any physical damage or obstructions	6	11	19
Ignore and try to reach my destination safely	5	8	8

In response to the third question about participants' potential reactions to detected anomalies, some participants reported that they would consult the vehicle manual. This reflects drivers’ reliance

on readily accessible in-vehicle resources to diagnose issues. It also highlights the importance of providing clear and easy-to-follow diagnostic information and guidance within the vehicle interface. Many participants chose to deactivate the affected system immediately and, if possible, engage manual control (6 in Scenario 1, 15 in Scenario 2, 31 in Scenario 3), with some suggesting a system reset might resolve the issue, drawing on familiar IT troubleshooting habits from everyday life. As a final step, several participants said they would report the issues to the vehicle manufacturer or an authorized service team, underscoring their trust in professional interventions for security and functionality recommendations.

## **5.2 Discussions**

### **5.2.1 Validation of Study 1 Hypotheses**

#### **5.2.1.1 In-Vehicle Drivers' Gaps in Handling Cyber Threats**

The anomaly detection accuracy comparisons suggest a significant discrepancy in participants' perception of anomalies across explicit versus silent attacks, thus supporting hypothesis H1. Specifically, explicit attacks differ from silent ones in displaying new objects or noticeable cues on the console display. Participants' awareness of the two explicit scenarios also differs. Scenario 3 was more easily detected due to salient vehicle motion changes (e.g., lane changing), implying that cyber threats causing motion changes induce higher driver awareness. Compared to Scenario 2, a higher proportion of participants in Scenario 3 chose to pull over, likely stemming from the perceived severity of the potential risks not only to those within the vehicle but also to external parties and the traffic system. Conversely, Scenario 2 led to fewer immediate pull-over decisions, with participants opting to observe for recurrence or possible troubleshooting.

Regardless of participants' awareness of anomalies, the accuracy in participants' understanding and diagnosing the three attacks remains limited, given their cybersecurity knowledge and the information provided within the vehicle. This finding aligns with previous non-driving studies, which have shown that only a minority of individuals can correctly handle fewer than half of the cybersecurity scenarios presented (Yan et al., 2018). These findings again question the general drivers' ability to identify and handle cyber threats within vehicles, even among active and experienced drivers who are more aware of cybersecurity concerns in driving systems.

### 5.2.1.2 In-Vehicle Display's Effectiveness in Drivers' Cyber Attack Handling

Analysis of participant evaluations suggests a neutral perceived effectiveness of the console displays in cyber threat detection and enhancing CSA across all scenarios, with Scenario 3 identified as slightly more useful yet not significant. Thus, the result partially challenges the latter part of Hypothesis 2, suggesting that the in-vehicle console display was not considered an impediment due to insufficient information provision, nor was it regarded as a contributing factor to participants' low accuracy in detecting and identifying cyber threats. Participants did, however, report that the display should provide proper alerts and guidance to support driver responses to the threats (see Table 5), particularly in the silent scenarios.

This finding suggests that drivers primarily relied on and trusted the vehicle system to detect and signal threats, rather than actively interpreting the clustered information to diagnose them. This may indicate the general driver's low willingness to delve into the system's underlying mechanisms to respond to cyber threats. Similarly, previous works raised concerns about potential distractions from additional information added in a driving environment (Payre et al., 2023), as the driving task is highly time-sensitive, and interpreting extra information will consume valuable response time. Given these findings, the in-vehicle information cluster displays should prioritize guiding users' subsequent actions over presenting additional detailed information. This was also supported by some participants' suggestions for improving the in-vehicle display design to better support cybersecurity response.

### 5.2.1.3 Potential Risk from In-Vehicle Responses to Cyber Threats

Hypothesis 3 receives support: drivers' inadequate comprehension of cyberattacks can lead to inappropriate responses, potentially increasing driving safety risks. Participants' inadequate awareness of cyberattacks reflected a general lack of cybersecurity knowledge and experience, subjective assumptions about threat patterns and severity, over-reliance on vehicle systems' capabilities, and abrupt decisions to disengage from driving. These factors may contribute to heightened risks under varying cyberattack conditions.

Most participants claimed they would directly opt to pull over if they detected unfamiliar anomalies, but the appropriateness of this response remains questionable in some conditions. For instance, executing a safe pull-over on a busy highway during peak traffic can be tricky. Moreover, pulling over might inadvertently catalyze additional social engineering attacks. Some participants reported that they would continue driving while ignoring potential anomalies. Their decision to pull

over depended on whether the anomalies would recur. Participants mentioned their experience with conventional system failures and local computer glitches informed this response tactic. However, this response might inadvertently heighten cybersecurity risks as continuing driving could unwittingly extend the time window for further malicious exploitation.

In-vehicle inspections and onboard diagnostic analyses were mentioned in some participants' anticipated responses. This response ties back to the concerns about the in-vehicle information display being distracting by providing extra information for diagnostics. The need for in-vehicle real-time diagnostics of cyber threats appears less realistic and time-consuming due to drivers' general lack of experience and knowledge, which may conversely increase the risk of driving hazards resulting from both inaccurate judgments and delays in prompt reactions.

### **5.2.2 Challenges in Modeling In-Vehicle Human Responses**

Study 1 was conducted to examine drivers' in-vehicle responses to cyberattacks and to assess whether this setting supports effective cyber threat detection and mitigation, serving as a suitable task environment for applying the CWA-informed ACT-R-based model to simulate human responses. Specifically, the study investigated how various forms of CAV cyberattacks influence drivers' cognitive processes and decision-making. However, the findings raised questions about the effectiveness of the current in-vehicle human responses to cyberattacks as the optimal application scope for the cognitive model.

Drivers' responses were generally less analytical (See Figure 20: In Scenario 1, over 97% of participants were unaware of the anomaly; in Scenario 2, over 79%; and in Scenario 3, nearly half.), anomalies went unnoticed or failed to prompt drivers' analytical process. A considerable proportion of our participants (See Table 5: 25 participants in Scenario 1, 32 in Scenario 2, and 48 in Scenario 3) tended to react by either pulling over or continuing without intervention when anomalies were detected. Other expectations from our participants, such as real-time onboard diagnostics or communication with supportive sectors, are not fully supported by the current in-vehicle environment. With the limited response options available in the current in-vehicle operating environment, general drivers' responses often bypassed the detailed analytical processing that the enhanced model is designed to simulate.

Beyond the current in-vehicle environment constraints, the lack of analytic cognitive process in drivers' responses, reflected in the tendency to pull over or restart the system regardless of the

attack form, is also constrained by drivers' limited cybersecurity task knowledge and the time-critical nature of in-vehicle operations. Without specific cybersecurity expertise, the diagnostic information provided could also become a cognitive burden during driving, potentially delaying reaction times if not properly trained. Moreover, the isolated perspective of a single vehicle with constrained access to the broader communication layer may restrict its capability to convey a reliable and comprehensive assessment of ongoing cyber threats to the driver. While some drivers expressed interest in conducting on-board inspection of cyberattacks, the combined limitations of driver knowledge and the in-vehicle environment cast doubt on the feasibility and effectiveness of relying solely on the driver to diagnose and appropriately respond to cyberattacks.

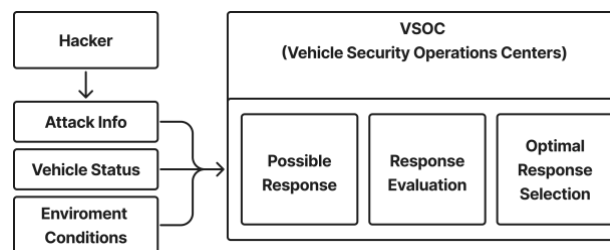
The proposed enhancement by integrating CWA insights into ACT-R based model is grounded in the assumption that the modelled task necessitates domain knowledge supported analytical cognitive processes, where human operators are equipped with sufficient cybersecurity knowledge to perform diagnostic analysis. However, in-vehicle human responses appear to deviate from this process, tending instead toward more immediate and direct perceptual-motor actions, even though these actions carry risks. This doubts the effectiveness of using the enhanced model to simulate driver responses within the vehicle's current environment.

### **5.2.3 Modeling Task Shift Toward Security Operations Centers (SOCs)**

In Study 1, participants also expressed expectations for guidance from authorized entities (e.g., vehicle manufacturers, authorized service centers) on appropriate responses to cyber threats, suggesting the need for external support for in-vehicle drivers. Given that CAV systems inherently rely on cross-sector collaboration, human roles in their cybersecurity defensive framework extend well beyond the in-vehicle environment, involving broader system-level coordination across legal, policy, and technical domains, reinforcing the collaborative efforts within the transportation ecosystem (Sadaf et al., 2023). For example, an emerging trend in CAV cybersecurity is the use of centralized data hubs that support real-time monitoring and coordinated responses from end-users to cyber threats, namely Vehicle Security Operations Centers (VSOCs) (Olt, 2019). The VSOC serves to bridge the gap between isolated end-user defensive strategies and the growing complexity of CAV cybersecurity (Gupta et al., 2023; J. Han et al., 2023; Katrakazas et al., 2020; Olt, 2019). J. Han et al. (2023) introduced VSOC for controlling and managing connected vehicle security resources and enabling driving guidance services in CAVs. Olt (2019) described having a connected vehicle SOC as

“combining security expertise with vehicle expertise.” VSOCs are still in the early stages of development. Recent studies have focused on optimizing their structured deployment within transportation systems (Martín-Pérez et al., 2023), advancing tool development (Saulaiman et al., 2025), and addressing the legal requirements for information sharing (Hofbauer et al., 2023).

The development of VSOCs as an extension of traditional SOC (Mutzenich et al., 2021; Olt, 2019; Sadaf et al., 2023). In many other complex socio-technical domains, centralized data hubs with defensive responsibilities are referred to as Security Operations Centers (SOCs), where a dedicated team of experts continuously monitors data to mitigate risks and prevent attacks, providing guidance, response, and evaluation to general end-users (Barletta et al., 2023; Hamad et al., 2024).



**Figure 21. The vehicle system shares attack information with the VSOC (Hamad et al., 2024)**

As a rapidly emerging critical component of cybersecurity operations, not only for CAVs but also across many complex socio-technical systems, modeling the task performance of SOC personnel can contribute to improving the design of defense automation tools, optimizing the allocation of SOC system functions, and informing overarching cybersecurity defense strategies (Rosso et al., 2022). Meanwhile, SOC analysts are expected to possess the necessary cybersecurity knowledge for effective monitoring, threat detection, and rapid response to incidents and events. In conclusion, the task environment of SOC analysts appears to be a more effective context for cybersecurity management, providing a suitable use case to demonstrate the model’s enhancements in both domain-specific analysis and the estimation of quantifiable measurements of human performance.

### 5.3 Limitations

Study 1 is a survey-based online study intended to capture broader attitudes and experiences. However, its validity is limited in reflecting real-world behaviors and consequences, as drivers may perceive and react to anomalous cues differently in a tangible driving environment. Therefore, the findings may not fully represent how in-vehicle drivers would behave in real cyberattack scenarios.

Moreover, as in-vehicle driving environment continues to evolve, the observed in-vehicle human behavioral patterns may not remain consistent over time. Therefore, future studies should keep pace with the driving system's developments by reconsidering the model's applicability to in-vehicle conditions, particularly if more information-processing and analytical decision makings begin to emerge in in-vehicle task operations.

This study does not directly present an ACT-R model; instead, it focuses on identifying an effective task environment for cyber-threat detection and mitigation to support future ACT-R implementation. By shifting from in-vehicle responses to more analytically demanding task environment as SOC alert triage, this work serves as a preparatory phase for validating and applying the model integration in practice.

## **5.4 Summary**

Study 1 was motivated by the need to identify a suitable task environment for modeling the effective detection and mitigation of cyber threats using the proposed cognitive model. It thus began by investigating in-vehicle responses to cyber threats, focusing on how different types of attacks with various display forms (explicit vs. implicit) and tampering with driving-related information (e.g., navigation, collision avoidance) disrupt drivers' attention and cognitive processing in detecting and responding to cyber threats. The Tesla Model 3 console display was used as a representative in-vehicle interface to examine how drivers perceive and interpret such attacks targeting the vehicle's information cluster.

However, the surveyed in-vehicle responses to cyberattacks largely reflected general drivers' limited awareness and consistently low accuracy in threat identification. First, most drivers failed to detect silent attacks. Second, even when explicit anomalies were noticed, drivers frequently struggled to develop an analysis to identify the threats. Although explicit forms of cyberattacks may raise drivers' awareness, most drivers' response did not engage in active analytical processing. This result indicated that in-vehicle drivers' responses to cyber threats were less effective, and generally lacking analytical decision-making processes. This ineffectiveness primarily stems from the limited availability of defensive resources of the in-vehicle environment and a general lack of cybersecurity knowledge among drivers, further compounded by the time-critical nature of the driving task. Thus, to the research question:

- *RQ2a. Is in-vehicle operation an effective task environment for detecting and mitigating cybersecurity threats?*

Study 1 indicated that current in-vehicle drivers' responses to cyber threats were less effective. And the current in-vehicle environment constraints naturally lead drivers to seek support from authoritative sectors within the broader cybersecurity defense framework, such as timely communication and salient alerts from external support sectors.

Accordingly, we look into the external cybersecurity support sector. Widely recognized as a best practice for securing assets, the SOC is increasingly being applied in CAV systems and across many complex and safety-critical domains. Its critical role in cybersecurity defense confirms it as an effective functional section in cyber threat management, requiring analysts to conduct appropriate incident analysis and provide prompt responses to events. So, to the research question:

- *RQ2b. Is a Security Operations Center an effective task environment for detecting and mitigating cybersecurity threats?*

The answer is yes. We therefore shift our modeling focus to the tasks of SOC, aligning with our observed end-user expectations (e.g., in-vehicle drivers) for external support with expert insights and actionable guidance to strengthen the system's overall security posture. In the next chapter, we will apply the proposed cognitive model to the SOC analysts' task, implementing it in practice and evaluating its enhancement within cybersecurity management applications.

## Chapter 6

### Study 2: Predicting SOC Analysts' Alert Triage Task Performance with a Cognitive Model Integrating CWA and ACT-R

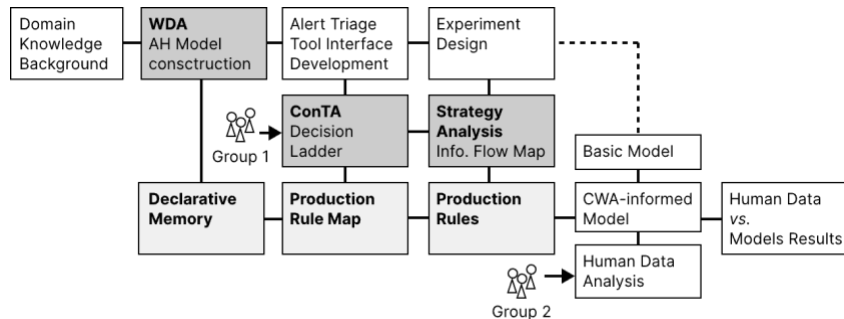
#### 6.1 Overview

This chapter builds on the findings of Study 1 by shifting the modeling scope from in-vehicle operations to broader cyber defense sectors' tasks, specifically, the performance of SOC (Security Operations Center) analysts' alert triaging task. This study will implement and apply the integrated cognitive model (see Chapter 4 for conceptual explanation) to simulate the SOC analyst's performance to address the third research question of this thesis:

- **Research Question 3:** *What are the enhancements, limitations, and future directions of integrating CWA and ACT-R for modeling human performance in complex domains like cybersecurity?*

This chapter is structured into two parts. First, we will follow the WDA to identify the fundamental functions of the SOC and build an experimental testbed for modeling analyst task performance. The second part elaborates on the detailed process of integrating CWA's domain-specific insights with the ACT-R-based model's construction and evaluate the outcomes of the integrated cognitive model. Finally, this chapter will discuss the enhancements of the integrated model, considerations for future applications, and limitations for further development.

As this chapter involves multiple stages, the following workflow offers a concise overview of its structure centered on the integration process and analysis.



**Figure 22.** Study 2 Overview: Integration Dimensions and Experimental Design.

## **6.2 Background**

### **6.2.1 Modeling Focus Shift from In-Vehicle Operations to SOC**

In this study, the modeling focus shifts from in-vehicle human operators' responses (see Studies 1 and 2) to SOC analysts. This shift may raise concerns about whether in-vehicle operations can currently serve as an effective cybersecurity defense role.

Given the design of the current in-vehicle systems, human drivers' primary goal remains safe driving (Zhang & Lintern, 2024), not cybersecurity defense. This finding is also reflected in Study 1's results.

However, this does not imply that in-vehicle operations are ineffective in cyber-threat defense. Drivers may be less aware of deceptive or stealthy threats without explicit alerts or system-level guidance (see Study 1), particularly given in-vehicle systems' limited access to correlated data for threat detection and mitigation (Hofbauer et al., 2023). Meanwhile, in-vehicle human operators remain the most affected and the direct observers of anomalous system behavior and contextual information. Their feedback and incident reporting can therefore play an essential role in supporting centralized cybersecurity monitoring, defense resource allocation, and strategies. In-vehicle operators remain a critical component of the defense framework.

Nonetheless, the current in-vehicle environment is not considered a primary defense context and provides limited support for effective cybersecurity detection and mitigation (Ding et al., 2025). As a result, modeling in-vehicle human operations may not offer the most realistic simulation of real-world cybersecurity human performance. Rather, SOC operations are increasingly adopted across many domains, including connected mobility systems (Hofbauer et al., 2023). Compared with in-vehicle response simulations, SOC analysts' performance modeling offers a more accurate reflection of real-world human performance within the general defense framework.

### **6.2.2 Internship Experience and Observations**

The author worked as a Human Performance and Human Factors Intern at CNL between 2023 and 2024, on a project that aimed to enhance the HMI design for operational technology (OT) SOC performance. The project was conducted in collaboration with two human factors scientists, four SOC subject matter experts (SMEs), and the author. The main technical SMEs team in the project included a senior cybersecurity engineer (technical lead), a computer scientist with a human factors

background (project manager), a senior SOC engineer (point of contact and coordinator), and a mid-level SOC analyst responsible for training new analysts and providing first-hand feedback on HMI design insights.

In the first year, the Human Performance Scientists and the author delivered an internal research report that included a literature review of SOC-related research, an iterative WDA of the SOC, and potential improvements to be informed by EID. This report was incorporated into the project's official Year 1 internal technical report (The technical report's literature review and the broader SOC WDA findings also inform the WDA presented in this chapter). In the second year, the author has contributed to the two rounds of focus groups interviews with the SOC SME team to further refine the functional allocation and task analysis of SOC analysts. The findings and outcomes from this phase were documented in the project's Year 2 technical report (The author's observations of SOC analysts operating, training, and insights into tool design improvement from SMEs help the experimental design of the alert triage task.).

Although the detailed research interview protocols, internal documents, training materials, and SMEs interviews are confidential, the literature review and the author's understanding and observations from these research activities reflect widely recognized concepts in SOC alert triage practice. These sources were used to inform the analysis and discussions throughout this chapter. For each dimension of analysis in this study, the specific methods and materials used are described in the corresponding sections, along with explanations of any confidentiality constraints (see Sections 6.4.2 and 6.4.9).

## **6.3 Introduction to Cybersecurity Operation Center Domain**

This section introduces the commonly accepted definitions and popular frameworks of SOCs. It provides a brief overview of SOC's core functions, existing challenges for SOC tools and preparation for the SOC analyst's modelling work.

### **6.3.1 Literature Review: Cybersecurity Operation Center Research**

#### **6.3.1.1 SOC Framework and Development**

Cybersecurity Operations Center (*SOC*, or *CSOC*) is widely recognized as the best practice for securing assets. SOC's core functions include "incident detection, analysis, and response" (Knerler et

al., 2023) and “protecting the information systems of an organization through proactive design and configuration, ongoing monitoring of system state, detection of unintended actions or undesirable state, and minimizing damage from unwanted effects” (*The Splunk SOAR Service*, 2023). Stewart and Jürjens (2017) defined SOC as a representation of the organizational security strategy combining processes, technologies, and people to manage and enhance the organization’s overall security posture. SOC’s definition is always with organizational role: “a centralized hub that operates at the heart of an organization’s network and security architecture and monitors an organization’s assets to detect, analyze, mitigate, and report security incidents” (Vielberth et al., 2020). While the evolution of SOC's advanced and tool-assisted functions varies, all definitions converge on a common SOC framework. This framework integrates “people, processes, and technologies” with key functions of “monitoring, detection, and mitigation” with its primary goals “operationally to maintain security and strategically to deter attacks” (Kersten et al., 2024.; Zhong et al., 2016).

Recent studies have examined general SOC frameworks, detailing the key functions, common workflows, and primary task analyses. Gamilla and Palaoag (2022) identified three distinct layers of a SOC: the *Core* layer encompasses the most active components including analysts, tools, and threat identification mechanisms; the *Policy* layer with a communication plan, incident response model and resources; and the *Standard* Layer comprised of recommend acceptability standards and serving as a mapping reference for the previous two. Shahjee and Ware (2022) suggested SOC consisting of a physical data source layer, a fault configuration administration performance security management layer, and an overall situation awareness layer. Essential features of effective SOC's usually include proactive threat mitigation strategies, native data feed integration, IT (i.e., Information / Communication Technology) and OT (i.e., Industrial Control Systems/Security systems) collaboration, automated threat intelligence feeds, highly trained personnel, and formal triaging procedures (Litherland et al., 2016; Parker et al., 2023; Takahashi, et al., 2011). In summary, SOC's are responsible for delivering situation awareness and data visibility to support their key functions, including monitoring, analysis, mitigation, alerting, threat hunting, and training.

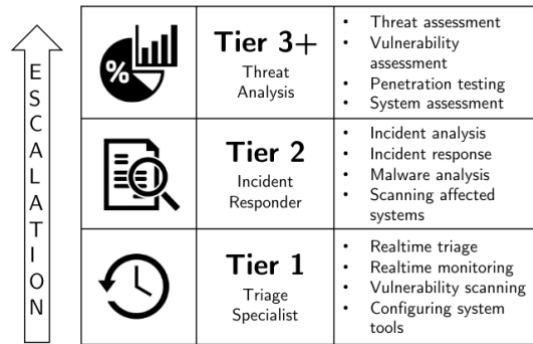
An automotive SOC has been similarly proposed as a clearinghouse for all security-relevant data from connected cars, enabling prompt analysis, risk minimization, and timely detection and mitigation of attacks (Olt, 2019). The database of an automotive SOC would primarily gather data from vehicles’ intrusion detection systems, including operational environments and threat analyses (Olt, 2019). This collection requires the timely and accurate capture of data from all potentially

exploitable systems and operational environments, across the broader connected automotive ecosystem to ensure accurate data correlation, threat detection, and effective responses. This aligns with the observation by (Li et al., 2023) that the real SOC environment, involving multiple business systems and assets, presents greater complexity in terms of discovery, evaluation, and analysis. The diversity of roles and expertise within SOC teams extends beyond domain knowledge of cybersecurity to encompass various domain-specific communications and organizational structures, significantly influencing SOC effectiveness (Petersen et al., 2020).

### 6.3.1.2 SOC Analysts and Expertise

The SOC teams typically include operators for specific response tasks and leaders for critical decision-making and coordination (Shahjee & Ware, 2022). Analysts at different levels collaboratively monitor network activity and respond to threats. Many studies agree that SOC analysts are hierarchically organized into three tiers (Agyepong et al., 2020; Eldardiry & Caldwell, 2015; Muniz et al., 2015; Takahashi, et al., 2011): (1) junior analysts coordinate monitoring, information filtering, isolation, and other routine tasks; (2) senior analysts handle vulnerability and risk management non-routine tasks and provide support to junior analysts; (3) managers engage in proactive planning, policy formulation, and process and project management. More briefly, SOCs are generally structured with Tier 1 analysts handling alert triage, Tier 2 and Tier 3 analysts focusing on incident response and threat hunting (Oniagbi, 2019).

Ideally, as shown in Figure 23, this three-tiered structure operates such that Tier 1 functions as the triage process within the detection stage of the NIST Computer Incident Handling Guide, involving real-time analysis of alerts, logs, and events (Kokulu et al., 2019). Tier 1 analysts filter out events and alerts, conduct simple research, and take actions based on predefined procedures or “runbooks” and tools such as SIEM to evaluate potential incidents (Forsberg, 2022; Knerler et al., 2023; Muniz et al., 2015). Complex cases or those lacking predefined procedures are escalated to higher levels, while critical incidents, especially, may be forwarded to IT forensics (Kokulu et al., 2019). As tier levels increase, analysts' responsibilities become more specialized, and the time required to resolve incidents tends to rise due to increased complexity.



**Figure 23.** Typical SOC analyst tier responsibilities (Kokulu et al., 2019).

### 6.3.1.3 Challenges for SOC Analysts

Prior studies have investigated the challenges faced by SOC analysts. Eldardiry and Caldwell (2015) highlighted the importance of sharing tools and effective communication among team members to optimize operations. Wang et al.(2021) noted limited research in behavioral and cognitive analysis for SOC performance enhancements. Ellis et al. (2022) emphasize the need for improved cybersecurity awareness of SOC with centralized dashboards featuring customizable interfaces, active and passive monitoring, and comprehensive data access. These studies emphasize the challenges posed by the information-intensive nature of the tasks and the need for supporting tools, which has led to growing research interest in improving SOC tool design. Werlinger et al. (2010) recommended enhanced tools and strategies to improve real-time threat response for challenges in accurate anomaly detection, managing cognitive workload, and minimizing false alarms. Kokulu et al. (2019) interviewed eighteen SOC personnel and collected critical issues contributing to the poor usability of current SOC systems, including low visibility of needed contextual information, information overload, and low-quality threat intelligence information. These challenges persist in today's evolved SOCs (Alahmadi et al., 2022).

### 6.3.1.4 Challenges in SOC Tools

Current SOC commercial tools, mainly Security Information and Event Management (SIEM) and Security Orchestration, Automation, and Response (SOAR), face challenges in processing diverse data ingestion and providing customizable data visualization to support real-time responses (Bridges et al., 2023). The primary function of SIEM tools is data collection and query. Comparatively, SOAR tools are designed to collect, filter, and present diverse information with extended capabilities by

automating repetitive tasks, enforcing standardized playbook-driven procedures, and enabling faster, more coordinated incident responses (Bridges et al., 2023).

Studies on the shortcomings of SOC tools have listed the following concerns:

- *Insufficient Information:* SOC analysts often spend additional time categorizing, correlating, and investigating relevant event details beyond these tools (Ranade et al., 2021). This is especially crucial for novice SOC analysts. For an inexperienced analyst, developing a complete and accurate understanding of a threat alert is complex, as it is easy to become overwhelmed or lose track of important relevant information (Kersten et al., 2024). While popular SOAR tools may increase operational efficiency, the lack of adequately configured domain-specific settings can hinder effective context-switching (Alahmadi et al., 2022), which hinders analytical reasoning during alert validation. Thus, the tools may reduce the accuracy and completeness of analysis outcomes (Kersten et al., 2024).
- *Overreliance on Automated Tool Indicators:* Instead of engaging in threat analysis, analysts tend to default to the suggestions provided by the system. Senior SOC analysts particularly emphasized the need to balance automation and human decision-making (Bridges et al., 2023). Comparatively, novice analysts mentioned the need for more explicit guidance and more detailed explanations of the displayed information (Kersten et al., 2024) to prevent over-relying on tools' intelligent indicators during threat investigations that are not always accurate or reliable (Kersten et al., 2024). On the other hand, the tendency to overlook tool-suggested low-severity alerts and the interference of redundant information on threat analysis are also among the top concerns for SOC tools (Karantzas & Patsakis, 2021).

These two main challenges in SOC tool development highlight the need to balance tool design between minimizing analysts' manual inspection of redundant data (Alves et al., 2021) and providing sufficient analytic support. This trade-off challenge, between reducing data overload and ensuring the delivery of actionable information, is where cognitive modeling, such as the proposed integration of CWA and ACT-R, could help in identifying practical information requirements by work domain constraints and human cognitive limits. Therefore, we consider implementing the integrated model and evaluating its potential to estimate human performance, which may inform future design improvements within this specific SOC work domain.

### 6.3.1.5 Alert Triaging in SOC

An SOC collects vast amounts of data from multiple sources, posing challenges for analysts in processing unfolding events and incidents (Salfati & Pease, 2022). Following the general SOC's 3-tier analyst framework (Andreassen et al., 2023), a Tier 1 analyst's main task is to use tools such as SIEM for monitoring and cyber threat alert triage (i.e., filter out events, conduct preliminary analysis, and take action based on predefined procedures or "runbooks" to prioritize alerts accordingly) (Knerler et al., 2022; Muniz et al., 2015; Forsberg, 2022).

In effective cybersecurity defense, alert triage, as the initial contact and assessment conducted by SOC experts, lays the groundwork for subsequent threat analysis, diagnosis, and timely response. While the triaging process generally follows structured pattern-recognition instructions, analysts frequently encounter unexpected situations due to the broad scope of alert sources, all while operating at a fast pace (J. Han et al., 2023; Khan et al., 2020; P. Sun et al., 2020).

The alert triage task is an illustrative example within the cybersecurity domain, as it is fast-paced operated within a dynamically complex work environment, and requires both domain-specific knowledge for routine operations and exploratory analysis. Accordingly, we will focus on the alert triage task for applying the proposed integrated cognitive modeling in this chapter.

## 6.4 Applying CWA in Modelling SOC Alert Triage

This section will start by applying three dimensions of CWA of the work environments and lay the groundwork for implementing and preparing ACT-R model constructions. The analysis includes two phases. The first part is the interface development for experiment preparation (from section 6.4.1 to section 6.4.6). The second part focuses on analyzing participants' alert triage operations on the developed interface, with modeling informed by their insights (from sections 6.4.10 to 6.5.8). Finally, the constructed model will be validated against the collected human data.

### 6.4.1 WDA and Abstraction Hierarchy

Work Domain Analysis (WDA), as the first phase of CWA, provides an in-depth interpretation of system functions and properties based on human reasoning patterns in complex systems (Vicente, 1999). WDA is comprised of two parts: a decomposition (or part-whole) hierarchy and an abstraction hierarchy (or means-ends) hierarchy (Vicente, 1999). Progression along the decomposition hierarchy results in a description of a different object (e.g., electric vehicle versus battery). In contrast,

progression along the AH's means-ends connections results in a new functional description of the same object (e.g., engine versus speed). The AH is rooted in Rasmussen's directions on identifying work domain constraints to articulate the system's objectives and constraints governing actions by examining human operator reasoning across different abstraction levels (Rasmussen, 1985). AH representations provide a way of representing complex work domains in a goal-directed problem-solving manner, especially in dealing with events that have not been anticipated (Vicente, 1999).

A systematic understanding informed by WDA of what SOC is and how it operates is necessary for modelling analyst performance and the effectiveness of SOC tools. Specifically, addressing current SOC tools' usability and functionality issues (Alahmadi et al., 2022; Kokulu et al., 2019; Oniagbi, 2019; Vielberth et al., 2020) requires a detailed understanding of SOC operations and the cognitive workflow bottlenecks analysts face. The outcome from the WDA also seeks to identify SOC's basic information requirements and offer insights to inform its tools interface design, and the preparation of declarative knowledge derived from WDA results for use in the ACT-R model.

In the following implementation sections, the domain analysis begins with a function analysis from the article review and the author's observations during the internship in Canadian Nuclear Laboratories (CNL) (see Section 6.2.2). The SOC functions are then mapped onto the five abstraction levels of the WDA tool's AH.

#### **6.4.2 Methods for Work Domain Analysis (WDA)**

The WDA of the domain-level functional constraints for a SOC Tier-1 analyst drew upon literature and document reviews, including a targeted literature review using keywords "security operation center", "task analysis in cybersecurity operation center", and "security operations function analysis" from peer-reviewed studies on SOC functional analysis and workflows. A total of 32 peer-reviewed selected research articles published from 2016 to 2023 were reviewed to establish the analytical foundation for the WDA. The literature review also includes CNL's internal design guidelines and additional documents from CNL's subject-matter experts' (SMEs) functional decomposition reports, which were accessed during my internship but are subject to the confidentiality agreement and thus does not reveal the details in this thesis.

To further refine and validate the analysis results, two rounds of collaborative design workshops were conducted with CNL's SMEs, during which the AH outputs were reviewed and iteratively improved. Two rounds of workshops were conducted with one human factors scientist, the

author, and four SMEs. The SMEs included: a senior cybersecurity engineer who authored and reviewed SOC functional-decomposition reports; a technical project lead with human-factors expertise; a technical director with the domain-specific cybersecurity experience; and a front-line SOC engineer with hands-on training and operational experience.

In the first workshop round, the human factors research team introduced the WDA framework and discussed its relevance to improving analysts' tool design. The distinction between functional analysis and WDA was clarified. The initial draft of the AH, developed from SMEs' technical documents, functional analysis reports, and relevant literature, was presented and reviewed collaboratively. During the walkthrough, any disagreements or uncertainties regarding AH elements were discussed in detail. When needed, SMEs provided practical scenario examples to refine interpretations and confirm the updates. The AH's revisions continued until consensus was reached.

Following the first round of workshops, a draft report including the revised AH was circulated to the four SMEs and the human-factors team for review. They were invited to provide comments, raise additional considerations, or suggest content informed by further reflection or related articles upon the first-round workshops' discussions. Two weeks later, a second round of workshops was held to address the newly raised comments and insights of the AH. Discussions led to a set of agreed-upon revisions. A subsequently updated version of the AH and workshop report was then shared with the SMEs and finalized after the second round of workshop review.

Because study 2 focuses specifically on Tier-1 analysts' alert-triage activities, the WDA in this chapter captures only the elements of Tier-1 analysts (It thus does not disclose any sensitive or confidential domain-specific functional elements from my internship).

### **6.4.3 Settings and AH Modelling**

The functional elements of SOC were derived from a review of articles review on SOC-related topics and supplemented by the author's observations during the CNL internship, which focused on a SOC tool design project and operational challenges.

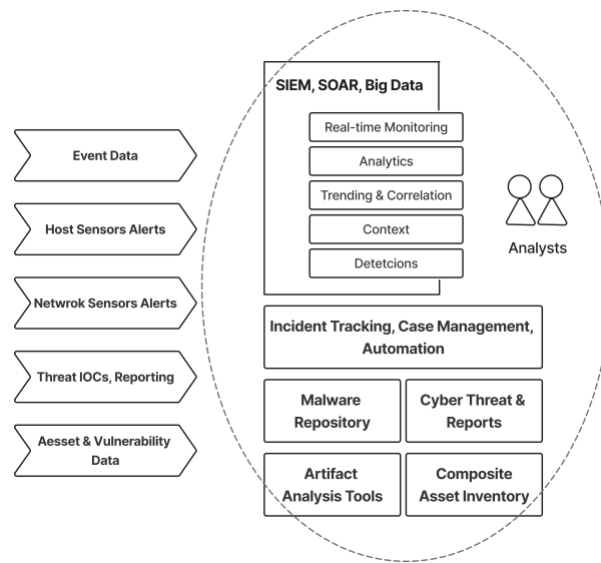
## **6.4.4 Conducting WDA**

### **6.4.4.1 Step 0: SOC operation basics:**

The essential missions of SOC operations include monitoring, detection, analysis, response, and recovery from all cyberattacks (Knerler et al., 2023).

Every SOC employs a range of technologies and automation processes to generate, collect, enrich, analyze, store, and present security-relevant data to responsible SOC members. Data sources for a general SOC include host sensors, network traffic metadata, and various log sources, such as application or operating system (OS) logs from devices, the cloud, or operational technology (OT). These sensors detect malicious or unwanted activity that warrants further attention from a SOC analyst. Beyond its own sensing capabilities, a SOC receives valuable information through various sources, such as email messages, phone calls, real-time text chats, walk-in reports, incident reporting forms on websites, and tips from other SOCs. Combined with security audit logs and other data feeds, this data is sent to automated tools for SOC, such as SIEM, to generate cyber threat intelligence.

Together, the cyber threat intelligence, security-relevant events from constituency assets, and organizational or system-specific information are fed into SOC operations. These inputs are filtered and assessed by both humans and automated tools to determine whether a response action is needed. Throughout this process, the SOC coordinates and consults with various stakeholders, such as system administrators and service owners, to ensure that any response actions taken align with the domain-specific environment the SOC supports.



**Figure 24.** Typical SOC data and tools (Knerler et al., 2023).

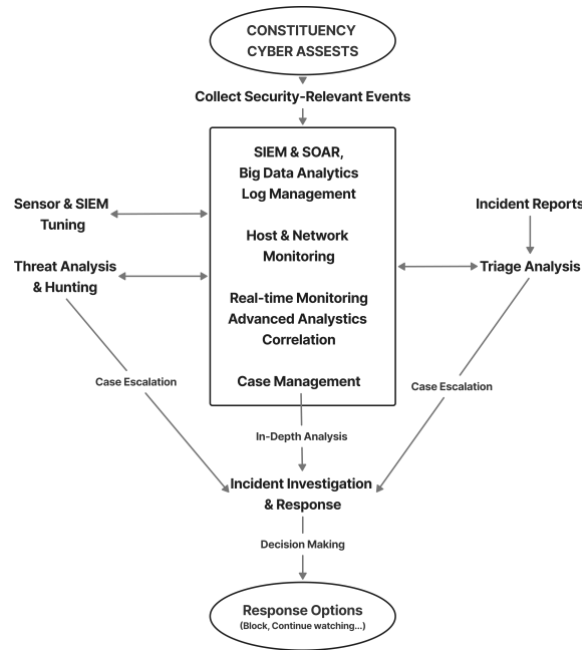
#### 6.4.4.2 Step 1: Determine Domain Boundary and Depth of Analysis

Conducting a WDA requires first identifying the analysis scope by establishing boundaries defining the targeted work domain (C. M. Burns & Hajdukiewicz, 2017). In this analysis, the core subsystem of focus is the alert triaging phase of SOC analysts' operations. Alert triaging is the first critical phase for human analysts in processing cyber-related data, linking subsequent stages of SOC operations such as response, eradication, containment, and prevention. This phase involves critical human analytic operation and decision-making, but research is still far from fully integrating human cognition into these processes (Alahmadi et al., 2022).

Threat alerts for triaging typically fall into two categories: signature-based and anomaly-based detections (Knerler et al., 2023). Signature-based detection relies on prior knowledge to identify malicious behavior, often using indicators of compromise (IOCs). IOCs are forensic artifacts of intrusions found on constituency systems at the host or network level, including details such as IP addresses, hashes, or malware characteristics. Whereas, Anomaly-based detection identifies deviations from normal behavior, triggering alerts whenever activity falls outside the expected deviation range.

In a typical SOC alert triaging workflow, an alert is triggered when a predefined set of criteria is met. Human analysts, supported by specialized cyber intelligence tools and automation, play a key

role in interpreting these alerts to determine whether further action (e.g., escalation) is required. Alerts must be evaluated within the context of the system(s) where they occurred, the surrounding environment, the supported mission, relevant cyber intelligence data, and other sources that can confirm and validate the cause and level of concern.



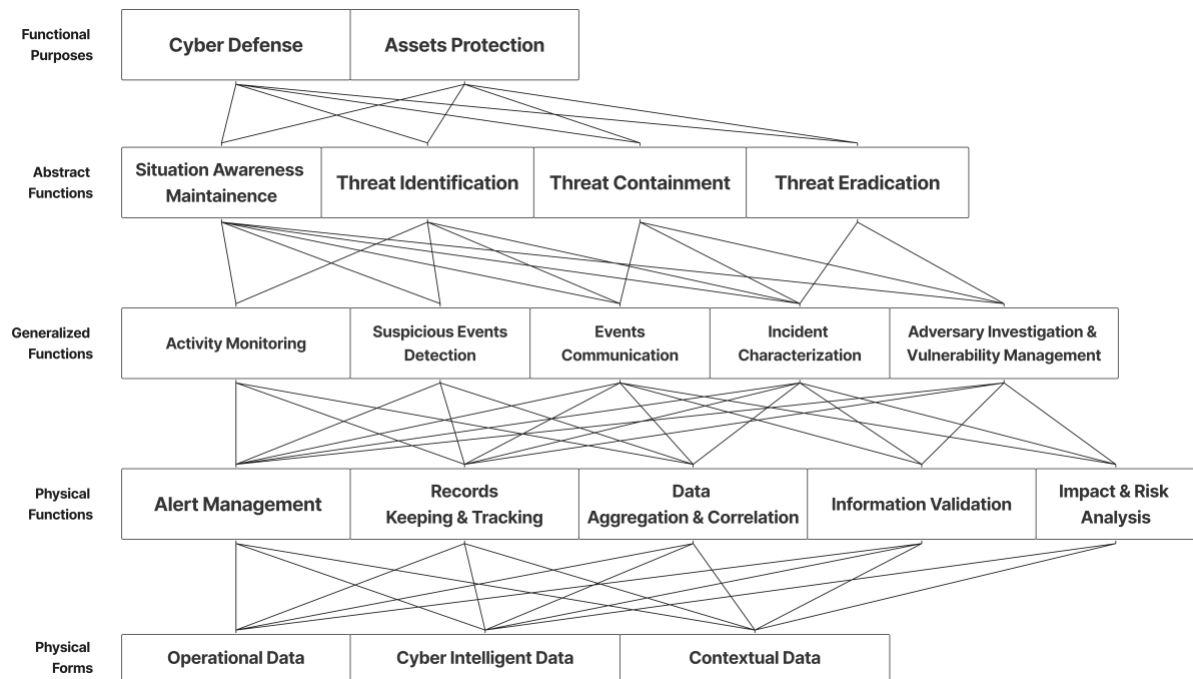
**Figure 25.** Basic SOC workflow (Knerler et al., 2023).

#### 6.4.4.3 Functional Purposes

A SOC can take on many different missions, but this study’s objective was to abstract a knowledge map for SOC’s alert triage operation environment. Hence, the AH begins with functional purposes, at the top of the hierarchy, which describe primary systemwide purposes. Two objects were identified: “Cyber Defense” and “Assets Protection”. These are the aim of establishing a SOC in many definitions (Agyepong et al., 2020; Knerler et al., 2023; Ofte & Katsikas, 2023).

#### 6.4.4.4 Abstract Functions

At the second level, Abstract Functions describe the principles that govern system operation (Burns et al., 2005). These principles support the SOC’s overarching cyber defense and asset protection goals. Key abstract functions include situation awareness maintenance, threat identification, threat containment, and threat eradication.



**Figure 26.** The AH work domain model for a SOC (on Tier-1 analyst work scope). The links illustrate means-ends relations between adjacent abstraction levels.

Situation Awareness maintenance involves sustaining awareness of active alerts, the organization's security posture, ongoing incidents, vulnerable assets, and current threat levels based on intelligence data (Newhouse et al., 2017). The corresponding tasks contributing to this function from the NIST Cybersecurity Workforce Framework (Newhouse et al., 2017) include T1047, T0291, T0469, and T0096. These codes refer to standardized task identifiers for categorizing and describing specific cybersecurity work tasks across roles and specialties. The following parts also use these task codes consistently to reference the same framework.

Threat identification refers to detecting and alerting potential security threats, which may or may not require further response. This element operates based on two commonly applied principles: anomaly detection and signature-based matching (Knerler et al., 2023). Anomaly detection involves establishing a baseline of normal or benign communication behavior between assets (Newhouse et al., 2017, T0023, T0994, T0977) and generating alerts when activity deviates from this baseline (Knerler et al., 2023). In contrast, signature-based detection relies on prior knowledge of known malicious behaviors, such as matching IOCs, to identify threats (Knerler et al., 2023).

Both threat containment and eradication are response to threats. The difference is that containment protects the system from operating in a normal status and prevents threats from further spreading within and to connected assets. At the same time, eradication aims to eliminate the threat's impact across the system. Threat Containment uses network segmentation and endpoint protection (Knerler et al., 2023) to isolate assets from threat impacts. Eradication primarily relies on analysis and investigation of the incident threat pattern and attack pathways, with simulation testing of system architecture to recover the system to its normal status. Threat Eradication will replace a component, redeploy baseline configurations, remove malware, and restore backup (Newhouse et al., 2017, *T0360*) to eliminate the threat impact.

#### 6.4.4.5 Generalized Function

The third level, Generalized Functions, emphasizes the key processes executed and coordinated within the work domain, shifting the focus to how the Abstract Functions are operationalized (C. M. Burns & Hajdukiewicz, 2017). For instance, attention turns to the processes that monitor and track asset status and activity in maintaining situation awareness. These include alert generation and characterization processes, communication of detected suspicious events, and subsequent investigation to determine whether a response or vulnerability requires further escalation.

Threat response, including containment and eradication, relies on incident characterization using IOCs and impact assessments informed by known vulnerabilities and prior incident analyses. These response functions are also closely linked to effective communication and coordination—both within and across teams and tools—to support timely and efficient threat mitigation.

Threat identification, primarily conducted through anomaly detection and signature-based matching, involves analyzing alerts and validating potential threats based on network activity, traffic patterns, packet contents, and configuration changes. As one of the most critical functions within an SOC, the threat identification process spans nearly all physical functions, serving as the foundation for subsequent detection and response actions.

#### 6.4.4.6 Physical Functions

Physical functions represent more specific capabilities afforded by implementing the above processes (C. M. Burns & Hajdukiewicz, 2017). For example, *managing alerts* generated from integrated automated tools, *maintaining and updating logs and records* for internal communication, and

supporting deeper investigation or building references for future incident detection and characterization. *Aggregating and correlating* these data sources enables effective activity monitoring, improves event detection, and enhances the accuracy of incident detection and characterization. Lastly, the organized records and logs help *validate* detections and reduce false positives, with more accurate *assessments of impact and risk*, ultimately supporting informed decision-making on threats.

#### 6.4.4.7 Physical Forms

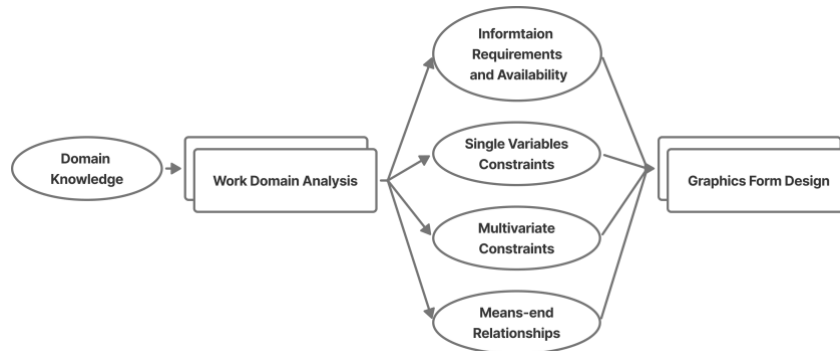
The last layer, Physical Form, describes the physical features of the components comprising the Physical Functions. SOCs accomplish the Physical Functions by gathering and curating extensive amounts of security-relevant data (Knerler et al., 2023). The Physical Forms level comprises three main categories of data: operational, cyber-intelligent, and contextual data. Operational data includes asset knowledge, log management, and event-based source data ingestion to SIEM for analytics. Cyber-intelligent data mainly includes IOC feeds, finished reporting, adversary and campaign knowledge base, indicator extraction, and detection and capture by automated analysis tools. Contextual data is more about the constituency technical environment and the business or mission context (e.g., adversary intent, resourcing, and interests toward an organization) (Knerler et al., 2023).

These data are where the SOC analyst understands the events, characterizes incidents, and locates the impacted assets for decision-making for responses. These data also provide essential knowledge and a detailed view for analysts to quickly identify asset impact scopes, configuration updates, and responsibility allocation for collaboration and communications for defense and remedy. The detailed contents of the three data categories are listed in Table 14.

#### 6.4.5 Translating to Interface Design Elements

Developing such a work model is the first step to transforming human cognitive paths into interface design solutions. The four-step design approach from EID details the development from the WDA model to the feasible and diagnosis-supported tool interface (C. M. Burns & Hajdukiewicz, 2017). Here, based on or constructed WDA model through AH, we translate the hierarchy into the information requirement (variables, constraints, and relationships) (C. M. Burns & Hajdukiewicz, 2017) derived from the model. The results of these information requirements (only variables) are summarized in Appendix B Table 14.

Following EID’s approach (see Figure 27), the next step is to establish the constraints and relationships for the information variables and inform the design forms accordingly. Although this part of the EID approach is not fully developed in this study, some simple single-variable constraints are based on a subset of the complete information requirements lists.



**Figure 27.** Systematic Approach to Graphic Form Design (C. M. Burns & Hajdukiewicz, 2017).

#### 6.4.6 Alert Triage Tool Interface Development

The interface used for this experiment is designed based on commonly used SOC tools (e.g., Splunk SOAR Phantom (*The Splunk SOAR Service*, 2023)). Figure 28 is a screenshot of its main alert dashboard. This study thus built a similar basic alert triaging dashboard using React.js as the standard interface toolset, replicating the general list view of alerts and a detailed information pop-up page. The alert information included in this experimental interface design is based on a subset of information requirements outlined in Table 14.

This experimental alert triaging tool includes two main pages:

- **Alert List** – displays all alerts in a structured format (see Figure 29).
- **Pop-up Detail** – provides additional alert details upon selection (see Figure 30).

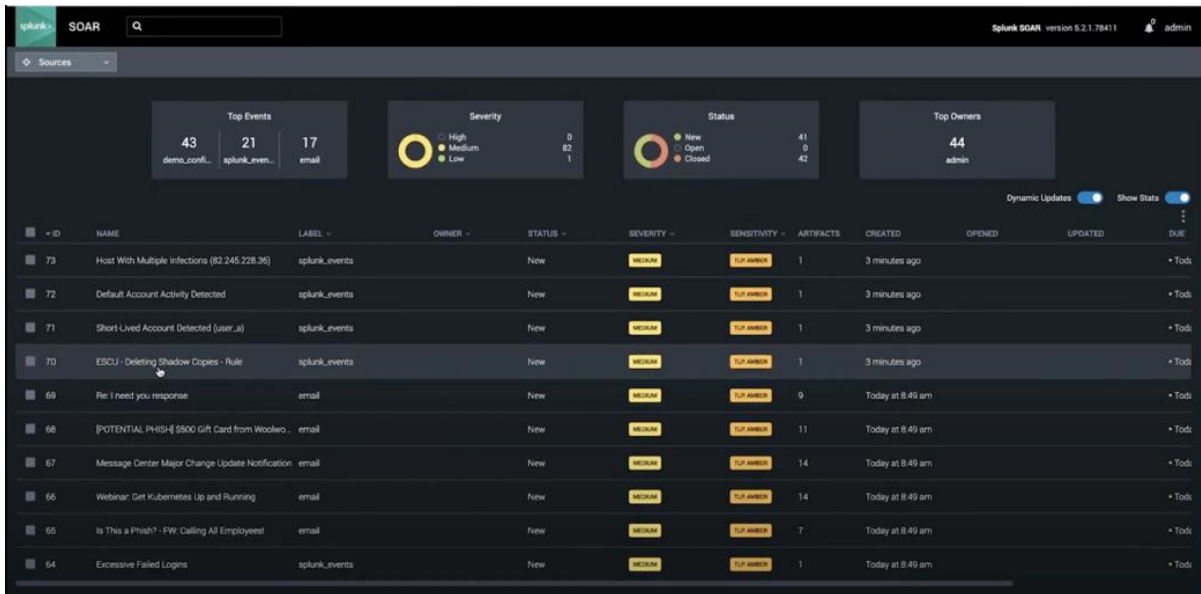


Figure 28. Splunk SOAR: Alert List view

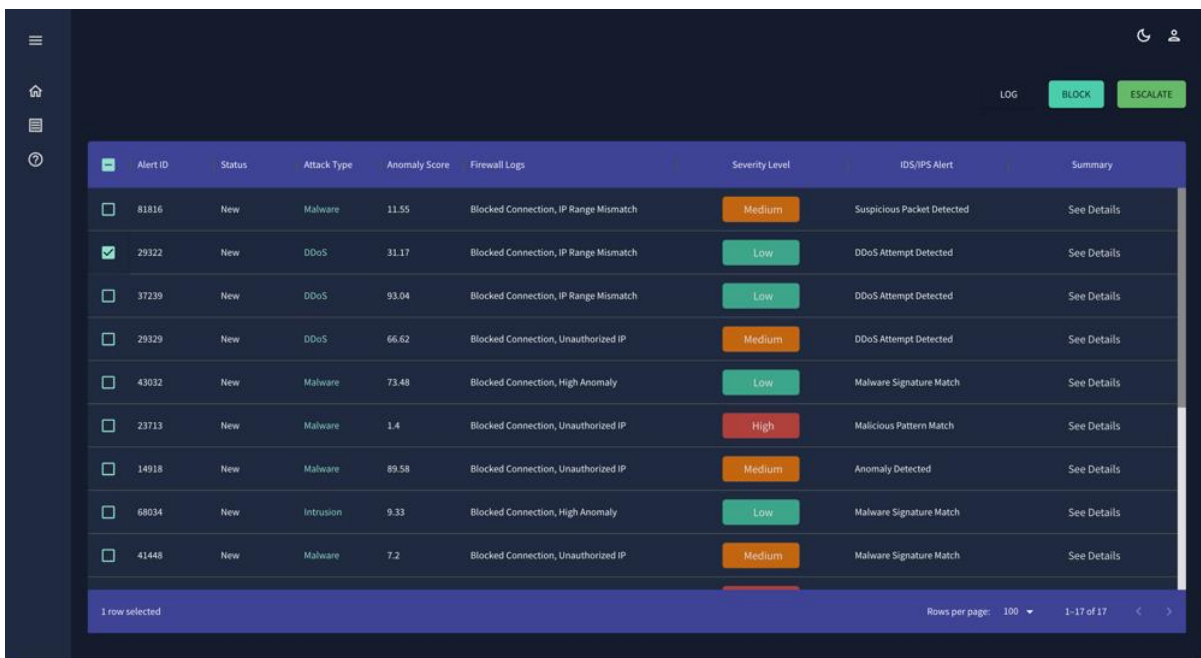
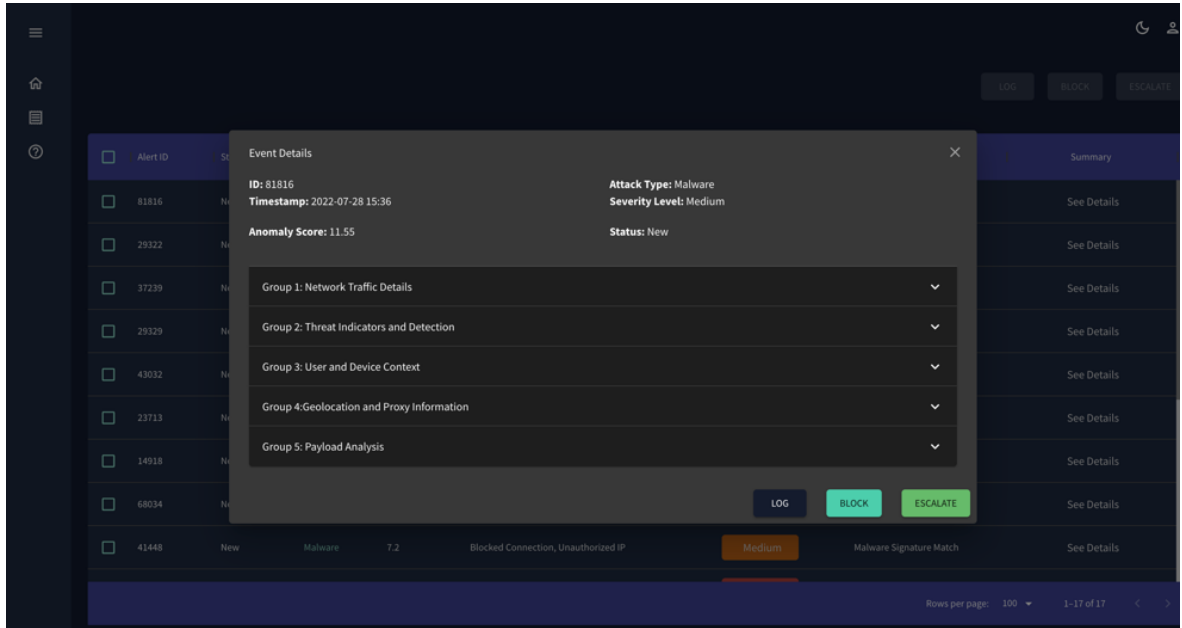


Figure 29. Experiment Alert Triage Tool: Alert-List View



**Figure 30.** Experiment Alert Triage Tool: Pop-up Detail

The information displayed in the interface was derived from the WDA (see Table 6). The identified invariants and variables can inform interface design following EID in Figure 27. However, the interface used in this study does not strictly follow the EID configuration, as the primary focus of this work is to explore the integration of CWA and ACT-R rather than interface design.

**Table 6.** The Mapping between a subset of WDA-derived information for Declarative Chunk Preparation (see the Full list in Appendix B Table 14).

AH Level	Variables Groups	Variables Displayed in the List View Page (Chunk Slot)	Variables Displayed in Details View
Physical Function	Impact & Risk Analysis	Severity level, Attack Type	
Physical Form	Operational Data	Firewall Logs	Network Traffic Details
	Cyber Intelligent Data	IDS/IPS Alert, Anomaly Score	Attack Signature, Malware Indicator
	Contextual Data		IP addresses, geolocation data, Road traffic conditions, Device Information, User information, Network Segment, Payload: command strings, malware payload

### 6.4.7 Experiment Design

This experiment followed a two-part mixed-methods design. The first part incorporated CWA dimensions to construct the integrated model. The second part collected empirical human-performance data. These data were then compared with the outputs of two constructed models: one informed by CWA analysis and one developed without. Pearson correlations and chi-square tests were used to compare alert-processing time estimates and decision-making alignment across models.

The overall procedure consists of two participant groups, one for model construction and the other for model assessment. Group 1 participants walked through the triage activities with the researchers. Their operations and interviews with researchers provided qualitative insights into analysts' response behaviors and decision-making strategies. These results were analyzed by the ConTA and Strategy Analysis frameworks. The analyzed outcomes were then used to build an ACT-R model that simulates Tier-1 analyst triage activity performance with the interface. A baseline ACT-R model without CWA analysis will also be developed to compare its accuracy and processing time with the CWA-informed model using human data. The real human performance data were collected from Group 2 participants for model results comparisons (see Figure 22).

#### 6.4.7.1 Alert Entry Data

The alert task dataset was sourced from the "Cyber Security Attacks" Dataset ([link](#)). The dataset was synthesized to provide a solid representation of cyber-attack vectors (e.g., attack signatures, threat types) as a good resource for alert analytical tasks. Although the dataset is highly comprehensive, this study uses only a small subset (17) of incident data as alert information for our tier-1 analyst triage task. The selection of alerts was assessed in a pilot study, which found that 5 participants completed 20 alerts in approximately 30 minutes, indicating a suitable online task time without fatigue or too much disengagement (Welz & Alfons, 2025). Thus, we limited the experimental workload to 17 alerts to maintain the quality of the results and ensure a balanced distribution of alert types.

#### 6.4.7.2 Participants

A total of 44 participants were recruited for the study.

Group 1 participants were recruited through the author's campus and professional network, targeting individuals with cybersecurity knowledge and/or prior experience working in a SOC environment or related areas (e.g., cybersecurity auditing and testing). The first group of participants'

data was primarily intended to develop the DL, Strategy Analysis, and was translated into Production Rules, and to provide subjective insights and experience into the alert triage process. Suggestions from Group 1 participants are also helping refine the task workflow and experiment design fidelity, making the task better reflect the work environment of SOC tier-1 analysts. This first group consists of 12 participants (see Appendix B, Table 17 for the participants list) (six had prior experience in cybersecurity or had worked as SOC analysts; two participants had cybersecurity knowledge acquired through formal education, past learning, degrees, or professional certifications (e.g., CISSP, CEH)).

The second group focused on collecting broader, objective behavioral data from a more diverse and general population, aiming to enhance the empirical foundation of the model's validation. A regression power analysis ( $\alpha = 0.05$ , power = 0.80) with one predictor and one outcome (i.e., task completion time) suggests that a sample of 40 participants is needed to detect moderate-to-large effects. Due to time constraints and the limited availability of expert participants, we expanded recruitment to include non-expert participants via the crowdsourcing platform Amazon Mechanical Turk. The second group recruited 32 participants, of whom 25 provided valid responses after excluding datasets with improper operations (e.g., batch processing without adequate observation time).

#### 6.4.7.3 Procedure

Group 1 participants were interviewed via Microsoft Teams video meeting after completing the 17 alerts triaging task using the developed tool interface. Before beginning, participants filled out a questionnaire to collect demographic information and assess their cybersecurity knowledge levels. They then walked through the runbook and instructions and completed five warm-up alert triaging practice tasks before the formal task. Completing the formal triage task, participants engaged in one-on-one post-task interviews with the researcher to explore the decision-making processes. They discussed their approach to deciding on different patterns of alerts. The interview questions are provided in Appendix B.

The second group followed the same procedure except for the post-task interviews. This included completing questionnaires, warm-up practices, reading the runbook and instructions, the formal task on the tool interface, and a final survey with open-ended questions, all in the same sequence as the first group participants. In particular, the behavioral traces of Group 2 participants were tracked and recorded for empirical data analysis.

#### **6.4.8 Defined vs. Ambiguous Pattern Alerts**

To enable rapid and automated responses in incident identification and detection, SOC teams usually rely on documented communication protocols, procedures, and general incident processing criteria outlined in a playbook (Applebaum et al., 2018; Tariq et al., 2025). A playbook is a document created by experienced SOC analysts that outlines the response steps and key alert indicators to follow and directs analysts' operations (Alahmadi et al., 2022; Jalalvand et al., 2025). The runbook used in our study defines a set of streamlined recommended indicator patterns, informed by analysis of the 'action taken' column in the original alert dataset and reviewed by an experienced cybersecurity analyst. Each pattern includes a set of features to inform decisions about specific alerts that match the pattern. Yet, this runbook's role is as a set of recommendations rather than prescriptive rules (Applebaum et al., 2018), and the final decision still depends on the participants' interpretation.

The alert data was thus categorized into two groups based on the difficulty of matching entry feature patterns with the runbook recommendations. The defined alert group (10 alerts) consisted of alerts that aligned with the feature patterns outlined in the instructions and runbook criteria (see Appendix B Table 16). The ambiguous pattern group (7 entries) consisted of alerts that did not match any of the recommended feature patterns, making them ambiguous. These ambiguous alerts required either further information inquiry or the participant's judgment to determine an appropriate response.

The alert patterns and relevant information identified in the runbook were made accessible to participants both in advance and concurrently via the interface website for on-demand reference (see Appendix B).

#### **6.4.9 Methods for Control Task Analysis (ConTA)**

In this study, the Decision Ladder (DL) was primarily derived from a literature review of Tier-1 analysts' responsibilities, workflows, and tool use. It was further supplemented with incident-response training materials for SOC analysts at CNL. Two rounds of focus group sessions with CNL SMEs were also conducted to examine analysts' tasks and activity flows during the second-year internship; the author participated in the sessions and transcribed the discussions.

The focus-group findings covered a broader scope of SOC activities beyond Tier-1 alert triage. The interview protocols and analysis have been documented in CNL's internal technical report but are not publicly accessible at this time. Therefore, this thesis focuses only on the Tier-1 alert

triage activity, as this component is generalizable across SOC applications in many domains and does not involve any sensitive or confidential information.

The DL was further refined using interviews from Group 1 participants, who described their decision-making processes and information-processing steps. The expert participants also provided feedback to enhance the fidelity of the testbed interface. The interviews were transcribed and analyzed using descriptive coding by two independent coders (i.e., the author and another research assistant) and reached consensus through discussion. The DL served as the organizing framework for mapping participants' decision-making patterns into corresponding phases (nodes and boxes) and for annotating the identified short paths.

These interview insights also informed the Strategy Analysis presented in Section 6.4.11.

#### **6.4.10 Decision Ladder (DL) of Tier-1 Analysts Alert Triage Task**

In the constructed DL, the triaging task begins when a tier-1 analyst notices a new alert (ALERT), prompting them to capture its details by analyzing the information from the cyber intelligence tools (OBSERVE). The analyst then assesses the alert's status to determine whether it is benign, such as an expected deviation within an acceptable threshold, or malicious (IDENTIFY). In some cases, the alert's status remains uncertain based on only a subset of indicators, necessitating additional observation, while in others, the alert is deemed urgent and requires immediate action.

Once the SYSTEM STATE is identified, the analyst must decide on the appropriate response. In the INTERPRET phase, the analyst evaluates the alert's possible impact based on either a predefined runbook—where meeting specific patterns may lead to a confirmed decision (GOAL)—or through their own judgment, relying on prior knowledge and experience (EVALUATE). This evaluation ensures that the decision aligns with the goals of the triaging (i.e., *Log* records the alert in the system with no further action required; *Block* isolates the threat and continues monitoring; *Escalate* refers the alert to a more experienced team member for advanced investigation.).

After reaching a decision (GOAL STATE), the next steps are planning and executing the appropriate response. The analyst defines the execution task (DEFINE TASK) and formulates the necessary actions (FORMULATE PROCEDURE) to complete the triage action.

Finally, in the EXECUTE phase, the analyst confirms the decision by clicking the appropriate action button (Log, Block, or Escalate) in the detailed information page or the list view.

Note that DL is never a sequential template for cognitive processes. Instead, it can enter at any point and transition between steps as needed. Shortcuts naturally occur, allowing for flexibility based on changing conditions and the user's experience.

Our analysis also revealed several shortcut patterns in DL. One example is the direct transition from ALERT to EXECUTE. This shortcut occurs when analysts initiate the triaging process by sorting alerts based on status or severity level rather than processing them in the default chronological order. Analysts may prioritize highly severe alerts first or begin with familiar alert types before addressing unfamiliar ones. Another is when an analyst quickly gathers deciding features and formalizes a response connecting the SET OF OBSERVATIONS and FORMULATE PROCEDURE. A typical case occurs when an alert has both high severity and high anomaly scores, leading some participants to escalate without further interpretation. This shortcut is particularly common when key indicators strongly align.

There is also a shortcut leaping between IDENTIFY and DEFINE TASKS. For example, expert analysts familiar with alert triaging tasks may skip in-depth interpretation and proceed directly without matching the predefined patterns in the runbook. In that case, the analyst can assess whether an alert meets predefined runbook patterns guidelines without explicitly recalling them.

A final shortcut category occurs when analysts revisit earlier stages after an initial interpretation or evaluation. If they determine their preliminary assessment is inconclusive or lacks sufficient supporting evidence, they may return from DEFINE TASK to OBSERVE to gather additional information before proceeding. This iterative process occurs when alerts share most feature patterns with typical alert patterns in the runbook but have missing or conflicting information, requiring further observation and assessment to ensure accurate triaging decisions. Another observed decision-making iteration is that, after processing several alerts, participants often revisit earlier decisions and revise them. This behavior may stem from cumulative observations and experiences gained as a result of processing more alerts. Participants may spontaneously iterate their assessments to be more stable and consistent if the initial attempts deviate too much, such as avoiding the imbalanced distribution of decisions (e.g., escalating too many alerts).

#### **6.4.11 Methods for Strategy Analysis**

Strategy Analysis was conducted through observation of group 1 participants' operations and their one-on-one semi-structured interviews, based on a protocol provided in the Appendix A and

supplemented by observation of their operational behaviors during task performance. The interviews captured participants' reasoning and decision strategies, while the observed operations could confirm what they described.

Two coders (the same as in Section 6.4.9) independently conducted descriptive coding of the Group 1 interviews with consensus-based discussions. Then the decision-making patterns were translated into input-output processes and categorized with the information flow map. Some

#### **6.4.12 Strategy Analysis to Information Flow Map**

While ConTA focuses on what needs to be done, Strategy Analysis details how it can be done. Therefore, we further examined these specific information-processing pathways by Strategy Analysis, drawing from observations and interviews with our Group 1 of participants.

Since the decision-making pathways are not fixed but vary depending on alert entry patterns and the operator's experiences, our strategy analysis aims to identify specific information and how it inform the triage decisions. The analysis also helps understand how the pathways switch and the factors influencing these switches.

Based on interviews with Group 1 participants, we identified and categorized three commonly used strategies, supported by illustrative quotes (The letter-number combination at the end of each quote refers to the coded participant identifier), as explained below:

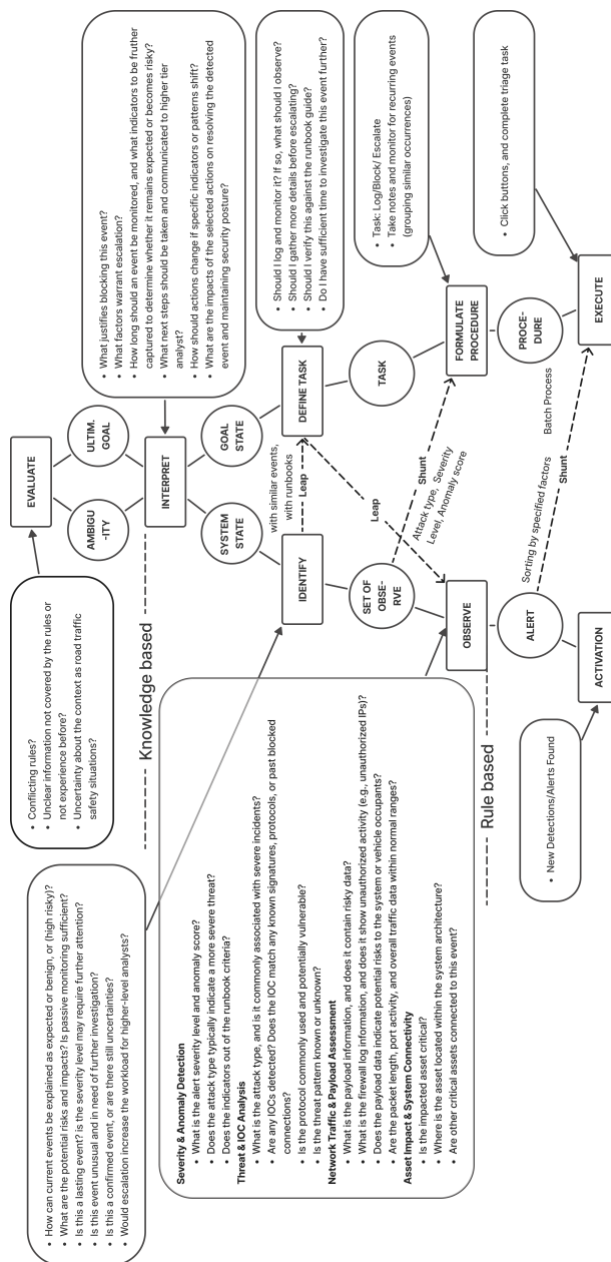


Figure 31. Decision Ladder for SOC Alert Triage Task

**Pattern-based strategy.** The first identified strategy is adhering to the runbook feature patterns. Since this strategy follows predefined alert feature patterns in the runbook, we refer to it as a pattern-based strategy. In this strategy, participants reference the key indicators listed in the runbook criteria (see Table 16), evaluate the alert's status, and take action accordingly. This strategy typically starts with identifying the attack type, followed by a sequential assessment of other indicators (e.g.,

severity level, anomaly score, firewall logs, and IDS/IPS information). However, path variations arise when specific features are absent, preventing the matching of the runbook recommendations. For example, suppose two out of three indicators match a recommended alert pattern decision from the runbook (In some cases, we have observed novice participants may still consider it a match and make a decision accordingly, even if other features are missing or misaligned).

*“Since the indicators didn’t fully align, I chose to be cautious and selected block or escalate... the severity level was marked high, while others, like anomaly score, were low. To be cautious, I escalated it for further review.” [P08]*

While checking the runbook criteria is the most structured and stable strategy, our observation indicate that it is not always entirely used, despite many participants claiming to base their decisions on the runbook and rated the runbook as useful (mean = 5.68 on a 7-point scale, where 7 indicates 'extremely useful'). Only 35.3% of (Group 2) participants actively referred to it at least once during their triaging tasks. This may result from matching each alert feature to the runbook patterns, imposing an extra cognitive workload of storing and retrieving alert patterns while making decisions. Therefore, analysts tend to switch to less cognitively demanding behaviors even though the pattern-based strategy is the most reliable. The tendency to deviate from a strictly reliable strategy suggests that, in real-world scenarios, analysts often balance structured guidelines with intuitive reasoning to optimize cognitive load and enhance efficiency.

**Streamlined strategy.** Building on the pattern-based strategy, participants also developed shortcuts, prioritizing visually prominent indicators that required less cognitive processing. Among these alert indicators, severity level, primarily displayed by color coding (see Figure 32), is the most immediately noticeable. Short texts and numerical values, such as anomaly scores and attack types, are also relatively easy to process. In contrast, detailed long-text descriptions from firewall logs or IDS/IPS require greater cognitive effort. From our observations, some participants only assess the severity level and cross-reference it with either the anomaly score or the attack type to make their decisions. For instance, an alert with a high severity level and its anomaly scores above 50 is generally identified as critical and is likely to be escalated. Specific attack types, such as DDoS or intrusion alerts with high severity, are almost always escalated.

In contrast, malware alerts with low severity are often considered less risky and are more likely to be logged. For long-text information in firewall logs and IDS/IPS descriptions, most

participants reported only focusing on key terms like 'Mismatch,' 'Unknown,' and 'Unauthorized' rather than reading the entire text. Considering the streamlined decision-making and assessment process, we refer to this strategy as the Streamlined Strategy.

*“Most of my decisions were based on severity level and the anomaly score.” [P04]*

*“My first consideration is always the severity level. If it’s high, I wouldn’t log it. At minimum, I’d block it.” [P10]*

*“... If (the firewall log) shows something like 'blocked connections' or 'unauthorized IP,' I’ll block it. If it’s already flagged as unauthorized, I will block it...” [P07]*

Alert ID	Status	Attack Type	Anomaly Score	Firewall Logs	Severity Level	IDS/IPS Alert	Summary
23713	New	Malware	1.4	Blocked Connection, Unauthorized IP	High	Malicious Pattern Match	See Details
14918	New	Malware	89.58	Blocked Connection, Unauthorized IP	Medium	Anomaly Detected	See Details
68034	New	Intrusion	9.33	Blocked Connection, High Anomaly	Low	Malware Signature Match	See Details
41448	New	Malware	7.2	Blocked Connection, Unauthorized IP	Medium	Malware Signature Match	See Details

**Figure 32.** A screenshot of Alert List view with main indicators

**Adaptive Strategy.** Adaptive Strategy. Expert participants demonstrated a more nuanced decision-making process by inquiring about a broader range of information, including asset lists, system topology, and additional contextual resources for verification. To solidify their assessments, experts usually ask detailed questions about the underlying meaning and context of indicators, such as:

*“Where is this information from, and is it trustworthy?” [P05]*

*“What is the difference between (“Severity Level” vs. “Abnormal score”)?” [P11]*

*“It would be helpful to have more details in the payload section. ... it helps to assess if there’s malicious content in the payload. This could provide more context for making decisions.” [P07]*

*“There should be an important link... a list of assets... Which assets will it affect?...to be intuitive enough to illustrates which asset it will affect ... it may have a greater criticality.” [P11]*

Compared to pattern-based and streamlined strategies, which rely primarily on indicators and information accessible on the list page, these experts' strategy involves exploring additional details

hidden in the pop-up page (see Figure 30), where decisions are made based on a broader set of features and information. We thus refer to this strategy as the *Adaptive Strategy*, which enables analysts to incorporate additional observations and contextual factors, rather than strictly following instructional recommendations, when making decisions.

Furthermore, experts usually develop their own assessment rules of indicators beyond the runbook instructions or system-generated indicators (e.g., severity level). Notably, analysts with the Adaptive strategy lack a consistent pattern. Rather than following a fixed set of paths, analysts utilize various types of information (see Figure 33) to refine their judgment and clarify the situation, ultimately making a final decision. Here, several quotes are presented to show key factors considered by expert participants in alert-related decision-making:

*“I mainly check packet length, traffic data, and port numbers. High packet lengths or unusual ports raise concerns. ...packets under four digits seem safer, while larger packets might pose risks like network congestion.” [P08]*

*“If the malware indicator is 'applicable' but the threat pattern is unknown, I'd escalate it as well. If both are known, I might log it.” [P10]*

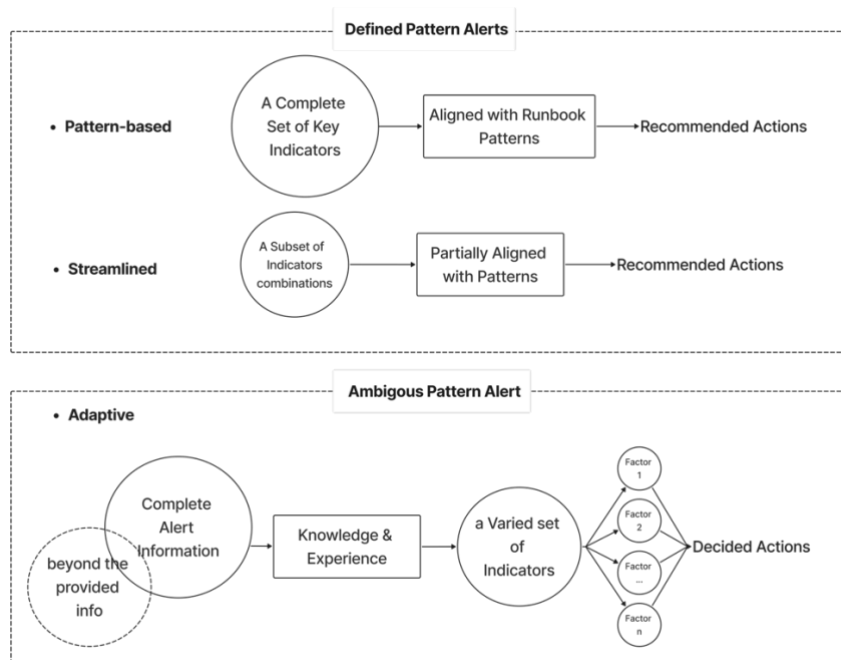
*“I focus on specific details, especially the payload information... I considered whether the attack could interfere with the vehicle's behavior, such as at a red light, causing it to move unexpectedly or become uncontrollable... If the vehicle is moving quickly, an attack would be more dangerous. However, in a slower scenario, like a red light, the risk seems lower.” [P12]*

Interestingly, one expert participant even factored in organizational considerations, including whether collaborators were outsourced and how this might impact access control and information confidentiality, and accounted for workload distribution [P11], pointed out that if too many escalations will affect the overall processing efficiency and ultimately influence his/her decision-making on this single alert. *“... the people behind the scenes (i.e., higher level analyst) will be busy ... and the frontline analyst's process will make no sense.” [P11]*

Thus, we find that these strategies are highly diverse and adaptive, not only from alert to alert but also across participants, especially with those alerts that do not match the runbook patterns (i.e., ambiguous pattern alert). Expert participants rely on a varied set of indicators (even beyond the

provided information) and draw upon their knowledge and experiences to make case-by-case decisions on those ambiguous pattern alerts.

The information flows representing the three strategy categories are: Pattern-Based, Streamlined and Adaptive strategies in Figure 33.



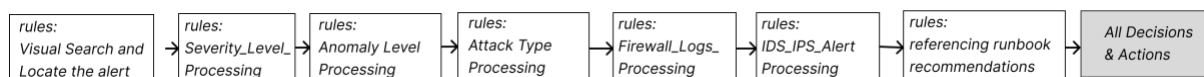
**Figure 33.** Pattern-based & Streamlined Strategies (Up); Adaptive Strategy (Bottom).

## 6.5 ACT-R Models Construction

### 6.5.1 The Basic Model

First, we constructed a basic model without further refinement by ConTA and Strategy Analysis, as our baseline. The model connects a set of basic visual stimulus processing rules for each feature displayed, as illustrated in the figure. These connected rules follow a single path to map the processed features to the corresponding feature pattern in the runbook to reach a final decision.

#### Basic Model Construction



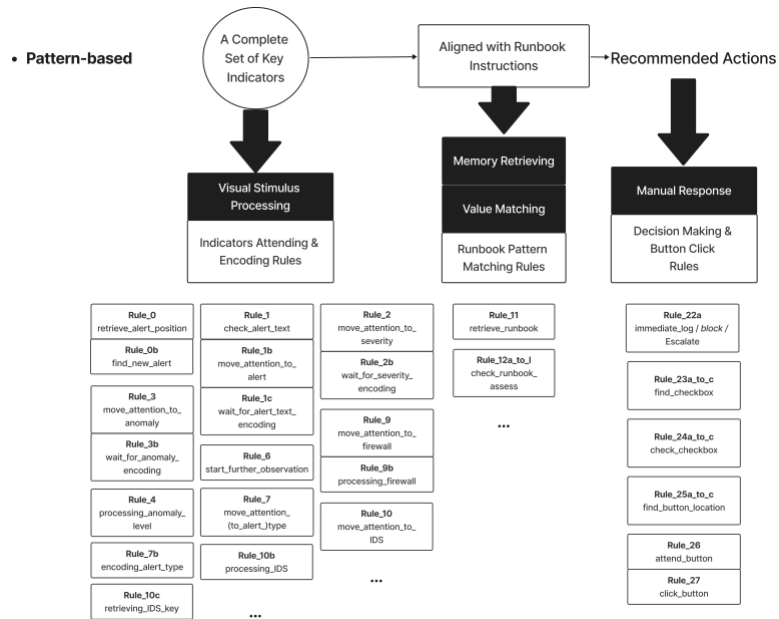
**Figure 34.** A Brief View of the Basic Model Rule Map.

The following section will explain how we translate the Strategy Analysis and ConTA results into the CWA-informed ACT-R model construction.

### 6.5.2 Strategy Analysis into Production Rules

The Strategy Analysis (see Figure 33) provides a brief overview of strategies used by participants as analysts to make triaging decisions.

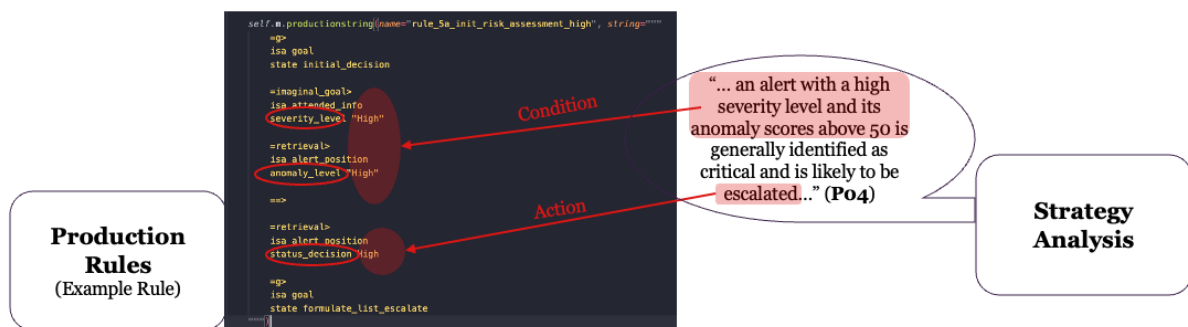
Among them, Pattern-based and Streamlined strategies are more structured and stable, whereas Adaptive strategies vary significantly in reasoning paths and necessitate more individual experiences and domain knowledge. We began by expanding the Pattern-based strategy as an example to build production rules. Following the information flow stages, each stage could generate a group of rules (see Figure 35). The production rules' IF-THEN structure serves as a microscopic tool, decomposing information flow stages into fundamental cognitive mechanisms, including attended stimuli, recall, value-matching evaluation, and motor execution as a set of rules. Notably, each set of rules in Figure 35 is only clustered for processing specific perceived elements (e.g., alert indicators, buttons), meaning that they are not connected to achieve the final goal in this stage yet.



**Figure 35.** Example of Pattern-Based Strategy Analysis and Corresponding Production Rules.

Next, the information flow from Strategy Analysis connects these feature-processing rules. For instance, the pattern-based strategy (see Figure 48 in Appendix B) must process more complete alert features before reaching the evaluation stage. Comparatively, the Streamlined Strategy forms shorter paths connecting fewer alert-processing rules (see Figure 49 in Appendix B).

For example, the production rules for processing severity level and anomaly score: when the two features are aligned, the strategy pathway leads to a decision based on their combined interpretation (see Figure 36). If both features indicate a high level, as shown in Figure 36, the resulting action is "Escalate". This brief example illustrates how production rules can be constructed to map specific feature conditions to corresponding actions upon Strategy Analysis. Similarly, more rule sets can be developed to process other combinations of features, leading to different evaluation and decision rule sets.



**Figure 36.** Example of Translating Strategy Analysis into a Production Rule Construction.

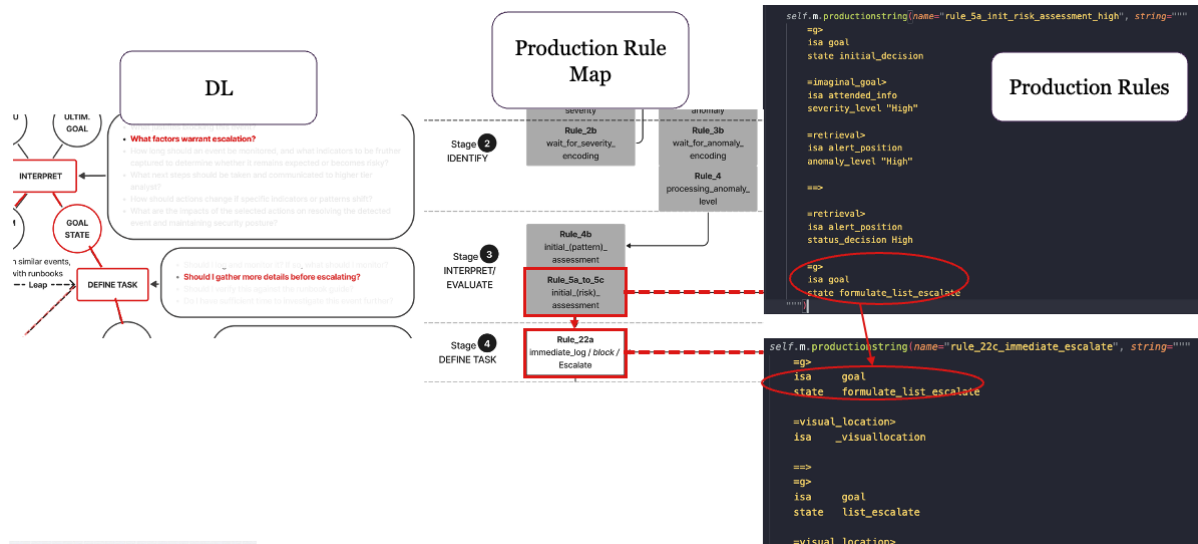
In conclusion, the Strategy Analysis could guide the construction of specific sets of production rules grounded in the perceived elements that require processing. These rule sets are established according to the information-processing pathways defined by each strategy.

### 6.5.3 Control Task Analysis to Production Map

In Figure 38, we map the distributed production rules derived from different strategies into the decision-ladder template with two considerations.

First, the DL maintains a consistent domain-general template that guides final goal achievement by connecting different cognitive phases through interactions with the task environment. It also helps identify which elements of the task environment must be processed to achieve the final task goal, using a recognized and cognitively plausible template. Unlike Strategy Analysis, DL

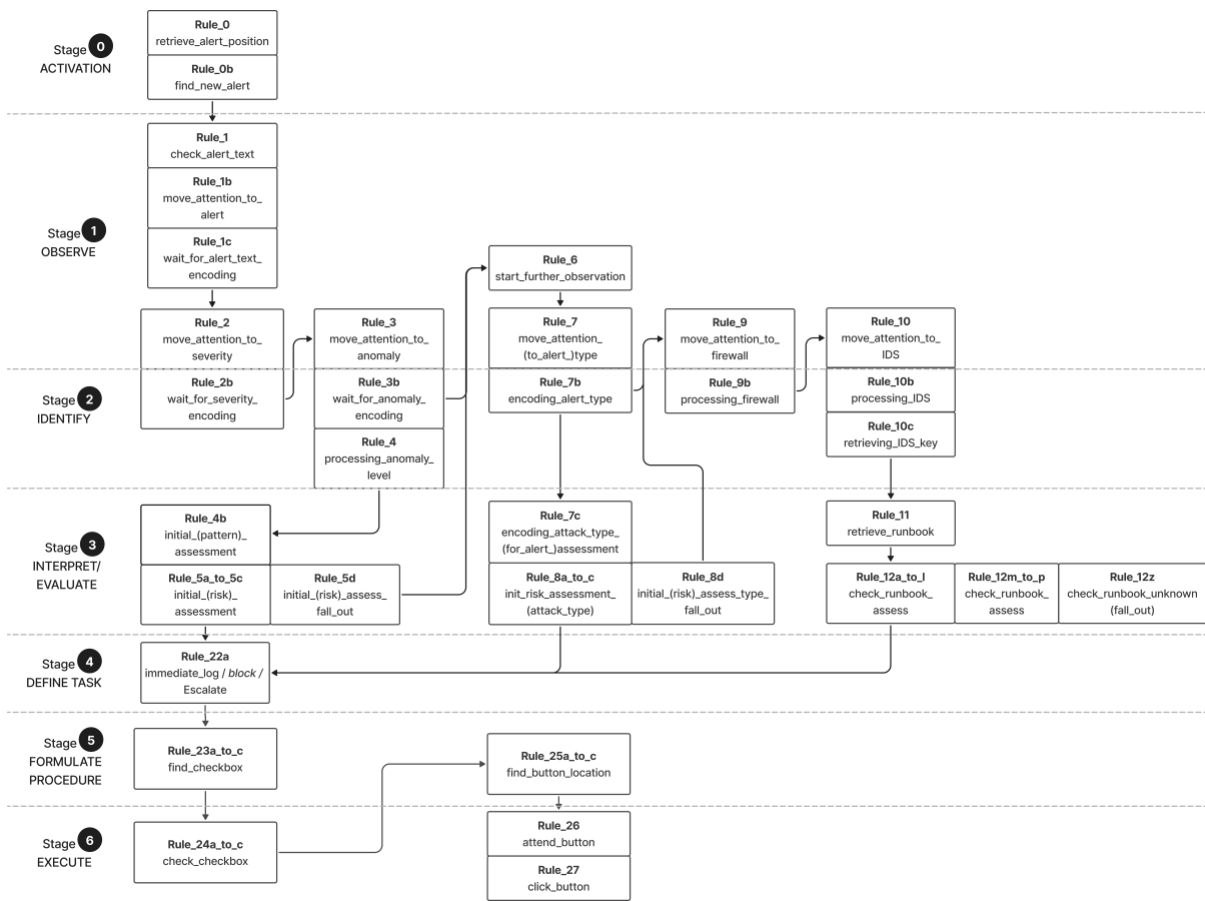
presents the full trajectory of human decision-making. As such, it completes the production rules map by introducing structural connections across the entire process.



**Figure 37. Example of how Production Rules are connected under the guidance of DL transition.**

This illustration comprises part of the DL (Figure 31), Production Rule Map (Figure 49), and the associated production-rule examples.

In practice, the DL supports the coordination and sequencing of rule sets derived from the Strategy Analysis by specifying how they can be connected and in what combinations. It also guides the selection of parallel pathways, strategy switching, and fallback configurations when conditions are not met (e.g., rule 5d, 8d, 12z in Table 18). For instance, the way production rules for processing alert features are connected can be determined using the established DL. Figure 37 illustrates an example in which two production rules (5a and 22a) are linked according to the DL transition from INTERPRET to DEFINE TASK. Rule 5a could be mapped to the INTERPRET stage by resolving “what factors warrant escalation?” The next activity, the DEFINE TASK, is to determine whether escalation is confirmed (output from rule 5 series) and whether to proceed by Rule 22a. This mapping from the DL to the production rule connection follows a coherent, recognized cognitive pathway for linking production rules. It thus makes the rule connections systematic and plausible in complex task environments.



**Figure 38.** Decision Ladder translated to the Production Map

The other consideration is that the DL also facilitates identifying missing connections and rules, or potential errors. As shown in Figure 38, rule sets that reside within the same DL phase tend to share a similar construction structure, making it easier to construct, organize, and check rules consistently within each phase. Unlike general production rule mapping, which reflects the model’s running procedure, the DL’s template offers a more interpretable cluster of rules that follows the human cognitive process. This enhances the usability of ACT-R models by making them easier to understand and communicate (DL phases indicate the specific type of activities each rule addresses), and adapt to future modifications (as rules can be replicated, added, or removed within the same phase of the DL).

A more detailed description of each production rule in Figure 38 and their mapping to the Decision Ladder is elaborated in Appendix B Table 18.

*Production rule* is the symbolic core component in the ACT-R model. ConTA and Strategy Analysis together provide structured frameworks to guide the production rules' construction and connections, integrating the strengths of both models into a more adaptable and systematic analysis approach.

#### 6.5.4 CWA to Declarative Modules

Before running the model with the above production rules, the declarative memory needs to be configured to support rule condition matching and knowledge retrieval. This model's declarative memory was populated with structured *alert\_position* chunks, representing the spatial and attributes of each alert entry. Each chunk included the *alert\_id*, and the screen coordinates of the key features (e.g., anomaly score, severity level, attack type, firewall logs, IDS/IPS alert) as in the list view.

**Table 7.** The declarative memory chunk (*alert\_position*), used by the model to store spatial and alert-related feature information.

Chunk Type	Slot
alert_position	<i>alert_id</i>
	<i>text x, text y</i>
	<i>anomaly x, anomaly y</i>
	<i>severity x, severity y</i>
	<i>attack type x, attack type y</i>
	<i>firewall logs x, firewall logs y</i>
	<i>IDS_IPS_x, IDS_IPS_y</i>

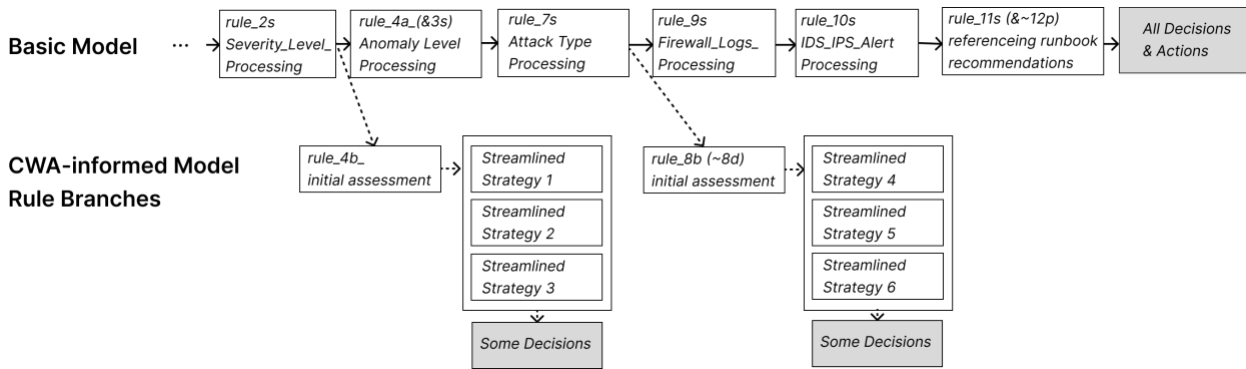
This chunk setup is a subset of the information requirements identified through the WDA, as this modelling scope is limited to list-view operations (see Table 6 for a selected subset, and refer to Table 14 for the complete list of information). The detailed reason for modelling only a subset of participants' operations and decision-making will be explained in Section 6.5.8.

#### 6.5.5 Overview of Differences: CWA-Informed Model vs. Basic Model

As the CWA-informed model construction is detailed above, here is an overview of the main differences between the two models shown in Table 8. As in Figure 39, we illustrate how the Strategy Analysis that informs the rules branching, demonstrates the overall structural differences between the two models.

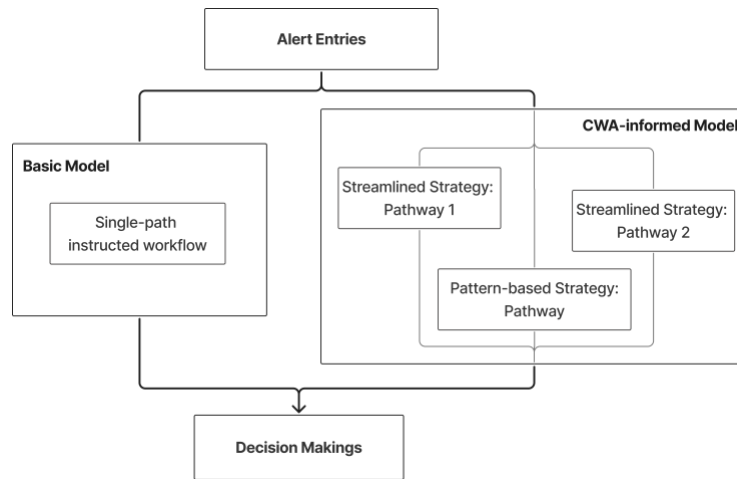
**Table 8.** An Overview of the Basic Model vs. the CWA-informed Model.

CWA Dimensions for the Model Construction	Basic	CWA-informed
<b>WDA:</b> declarative knowledge preparation	Yes	Yes
<b>ConTA:</b> feature selections based on different strategy paths, and the rule map connecting	Partially	Yes
<b>Strategy Analysis:</b> more flexible strategy paths for rules construction	No	Yes



**Figure 39.** Structural Comparison of the Two Models and Rule Branches in the CWA-Informed Model Based on Strategy Analysis.

More specifically, as shown in Figure 40, the key difference lies in the branching from the Strategy Analysis, which shapes the different pathways in the task model structures of the basic and CWA-informed ones. For instance, the basic model follows a single, direct path that processes all alert features according to the runbook’s recommended criteria and then reaches a final decision by connecting production rules for each feature processing. In contrast, the CWA-informed model allows multiple possible pathways under the same conditions. So, the model enables deviations from the Pattern-based Strategy to Streamline Strategies (e.g., Pathway 1: Rules 5a–5c; Pathway 2: Rules 8a–8c, as in Table 18. Production Rules Descriptions) through fallback rules (e.g., rule 4b, rule 7c) transitions.



**Figure 40. Differences in production rule connections between the Basic Model and the CWA-informed Model.**

(Streamlined Pathway 1 includes rules 5a to 5c; Pathway 2 also includes rules 8a to 8c, see Table 18.)

### 6.5.6 Global Parameters

The global parameters for the model have been kept at their default values, except that we adjusted the *cycle\_time* parameter from the default of 0.05 to 0.25 seconds to account for task complexity. The *cycle\_time* refers to the fixed time interval between successive production rule evaluations, representing the architecture's internal processing resolution (Anderson et al., 2004). The cycle time adjustment is typically used to simulate age-related cognitive slowing (Jastrzemski & Charness, 2007; Salvucci et al., 2004); it was used here to capture task complexity-induced cognitive difficulty and task difficulty for novices. A similar effect could alternatively be achieved by decomposing existing rules into smaller sub-steps to simulate the increased task complexity in future model refinements.

### 6.5.7 Rewarding Mechanism and Utility Manipulation

Although this study primarily focuses on the symbolic part of the model by translating CWA analysis into ACT-R rules, the sub-symbolic component remains an integral part of the modeling process. *Utility* is a key component of ACT-R's sub-symbolic reinforcement learning reward mechanism, allowing dynamic adjustments to how rules are selected based on the changing utility.

$$U(p) = V(p) + \alpha \cdot i \sum (r_i - V(p))$$

Where  $U(p)$  represents the utility value of production rule  $p$ .  $V(p)$  denotes the average past reward associated with rule  $p$ .  $r_i$  is the reward value obtained from the  $i^{th}$  execution of the rule.  $\alpha$  represents the learning rate, which controls the speed of utility updates.

**Table 9.** Specific non-default Utility Assignment.

Rules	Utility	Probability*	Reasons for Assigning Non-Default Utility
8d_init_risk_assessment_fall_back	-1	17.7%	If the conditions of the other Rules (8a to c) are not met, this rule will serve as a fallback, firing when the alert pattern does not match the first round of assessment of the alert features
12m to p_check_runbook	-1	17.7%	Reflecting the Streamlined Strategies, which has a lower likelihood of firing when the Pattern-Based Strategy is effective (Rules 12a to l)
21z_check_runbook_unknown	-20	~0%	if the conditions of the other Rules (12m to p) are not met, this rule will serve as an overall fallback, firing when the pattern does not match any runbook criteria or strategies, the value of -20 discourage selection unless no better rule applies
* This column's values correspond to an expected firing probability for the assigned utility rules when competing with a rule with no specific utility value assigned (with a default noise parameter $\theta = 0.25$ , and the calculation could refer to (Bothell, 2004).			

In this effort, we modified certain rule *Utilities* to regulate rule switching under specific conditions (see Table 9). For example, some rules from different strategies operate in parallel (e.g., Rules 12a to 12l are from the Pattern-based Strategy, and Rules 8a to 8c, and 12m to 12p are from the Streamlined Strategy), meaning they can all potentially fire under the same conditions. However, the likelihood of firing a specific parallel rule depends on the constraints imposed by work conditions, expertise, time pressure, or fatigue status, among other factors. In running the model, this switching mechanism is controlled by utility assignments to regulate the probability of transitioning between different pathways. In particular, Rule 8d and 21z (see Appendix B Table 18) were assigned a below-average utility value (-1, -20) and served as a fallback rule when other conditions were not met, ensuring the continuous flow of rules within the model. We also assign a lower utility to the Streamlined Strategy rules when they are paralleled with Pattern-Based Strategy rules, reducing their likelihood of firing when the Pattern-Based Strategy is more effective in achieving better accuracy (see Rules 12m to 12p in Table 18).

### 6.5.8 Exclusion of Adaptive Strategy Modeling and Ambiguous Pattern Alerts

The Adaptive Strategy for an ambiguous alert was not modeled in this study. The first reason is the highly dispersed cognitive pathways in Adaptive Strategy, meaning that decision-making patterns vary from case to case. Still, we aim to avoid oversimplifying or generalizing the Adaptive Strategy, as it risks biasing cognitive path flexibility and contradicting the core intent of Strategy Analysis. Second, some factors considered in the Adaptive Strategy extend beyond the interface-displayed stimuli (see section on Adaptive Strategy. and are thus challenging to capture through rule-based modeling. Therefore, we chose not to represent the Adaptive Strategy for ambiguous alerts in this study.

## 6.6 Results

### 6.6.1 Participants' Behavioral Data Analysis

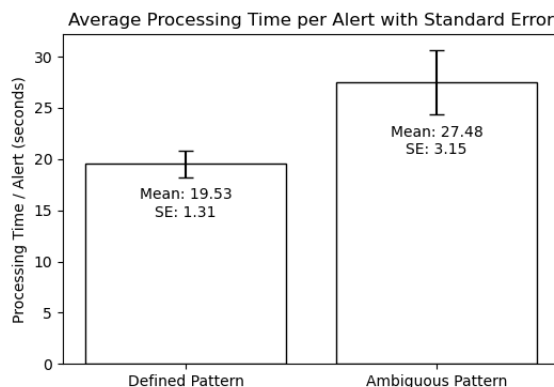
A total of 25 qualified participants (out of 32) in Group 2 triaged 10 defined-pattern alerts and seven ambiguous-pattern alerts. We hypothesize that participants use different decision-making strategies for these two categories, as we identified for Group 1, resulting in differences in their behavioral performance.

The average processing time for all 17 alerts was 12 minutes and 21 seconds (SD = 9 minutes 12 seconds), calculated as the time between a participant accessing the alert list and completing the final alert. Due to the high variability, we further analyzed per-alert response times to identify contributing factors to the variability in response times. Processing time per alert was calculated based on the following criteria:

- If alerts are processed individually, the processing time for each alert is measured as the time difference between the previous action and the timestamp of the current alert's action button click.
- If multiple entries are checked before clicking the same action button, the processing time for each alert is calculated as follows:
  - For *the first alert*: Processing time is measured from the timestamp of the previously actioned alert entry to when this alert's checkbox is checked.
  - For *the last alert*: Processing time is the difference between the timestamp of the previous checked alert and the button click.

- For intermediate alerts, Processing time is calculated as the time between checking the previous and current alerts.
- Only the first action is recorded when multiple actions are performed on the same alert (When the follow-up action doesn't occur sequentially, it becomes difficult to track its start time and measure processing duration).

A Wilcoxon signed-rank test compared per-alert processing times between defined and ambiguous pattern alerts. The results revealed a statistically significant difference between the two categories ( $W = 63.0, p < 0.01$ ), with a moderate effect size (Hedges'  $g = -0.44$ ). Figure 41 illustrates that the average per alert processing time was significantly shorter for defined pattern alerts (Mean = 19.53s, SE = 1.31s) compared to ambiguous pattern alerts (Mean = 27.48s, SE = 3.15s). Additionally, the more significant standard error (SE) observed for ambiguous pattern alerts suggests greater variability in processing times.



**Figure 41.** Per Alert Processing Time (Defined Pattern vs. Ambiguous Pattern)

Differences in per-alert processing time may be due to participants exploring alert features presented in different views. For example, when the alert was ambiguous, participants tended to click on the "See Details" option in the list view (see Figure 29) to access additional supplementary features and gain a better understanding of the alert conditions. Thus, we hypothesize that participants engage in more in-depth exploration for ambiguous alerts compared to defined ones, reflected in increased clicks on "See Details," as more frequent viewing of supplementary information, and ultimately longer processing times.

However, the Chi-square test results were not statistically significant ( $\chi^2 = 1.020, p = 0.60$ ), indicating no statistically significant relationship between alert types (Defined vs. Ambiguous) and

participants' clicks on "See Details." A closer examination of participants' behavior reveals that while accessing the details page, they rarely expanded the accordion sections containing additional grouped information (see Figure 30). This suggests that Group 2 participants may not effectively use the details view to actively explore supplementary information.

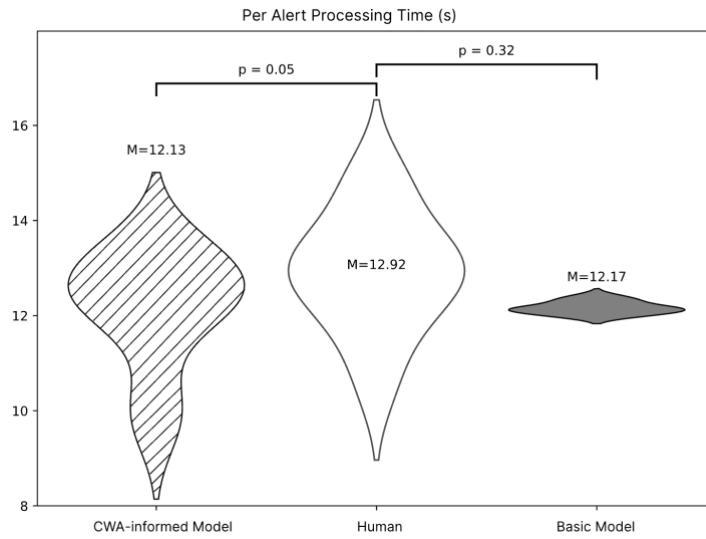
Considering such uncertainty in participants' use of the "See Details" page, we chose to exclude instances of defined-pattern alert processing that involved clicking into the details view. This reduced the dataset from 250 trials (25 participants  $\times$  10 defined-pattern alerts) to 161 trials used for the following model validation.

### **6.6.2 Human Data vs. Models Simulation**

We ran the constructed model using the Python-based ACT-R package (*pyactr*) and tested on 10 defined-pattern alerts, with interactions restricted to list-view operations only.

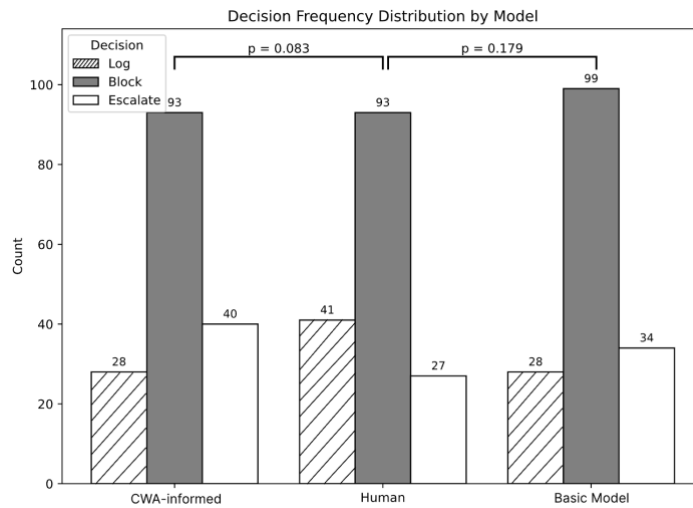
Prior ACT-R studies ran between 10 to 50 repetitions for stochastic simulations (Marewski & Mehlhorn, 2011; N. A. Taatgen & Lee, 2003). In this study, 25 simulation repetitions were performed for each alert entries in both models. The simulated results show that the average coefficient of variation (CV) (Reed et al., 2002) across CWA-informed model runs was 13.95% indicating moderate variability, and the basic model exhibited a much lower CV of 3.87% with high stability. The moderate variability observed in the CWA-informed model arises primarily from systematic structural differences (see Section 6.5.5 and Figure 40). Specifically, the variability in the CWA-informed model arises from the diverse information-processing pathways informed by the Strategy Analysis, rather than from random noise. This is also evidenced by the basic model's high stability across the same number of runs and task environment setups. Given both models' variabilities already converge after 25 runs to yield stable patterns, this number of runs is sufficient.

The overall average human per-alert processing time for defined-pattern alerts in the list view was 15.76 seconds ( $SD = 5.87$  seconds). We further removed outlier values defined as those exceeding  $\pm 2 SD$  from the average (Howell, 2013) (i.e., below 4s or above 28s), as too long or short durations raised concerns about the quality and reliability of participant engagement. The clean data average per-alert processing time comparison with the two models' results is illustrated in the figure below.



**Figure 42.** Human vs. Simulation Processing Time per Alert.

Pearson’s correlation between CWA informed model-predicted and human-observed per-alert processing times was  $r(8) = .62, p = .05$ , root mean squared error (RMSE) = 1.30 s, the basic model’s correlation test shows  $r(8) = .35, p = .32$ , RMSE = 1.39 s, neither showing significant effect in the estimating accuracy (see Figure 42).



**Figure 43.** Models Agreement with Human Decisions. (Each vertical bar reflects the frequency of a particular decision type produced by the corresponding model and human. Frequencies of decisions

are displayed above each bar, with horizontal lines linking model pairs that are subjected to Chi-square tests. Corresponding p-values are shown above the connectors.)

Figure 43 presents a grouped bar chart comparing the agreement of decision outcomes (Log, Block and Escalate) between the two models (simulation results were aligned with the human dataset by correspondingly excluding missing data): the CWA-informed model vs. Human data, and the Baseline (non-CWA) vs. Human data. Neither model exhibited a statistically significant difference from human decisions based on the Chi-square test (CWA-informed model:  $\chi^2(2) = 4.97, p = 0.08$ ), the baseline model ( $\chi^2(2) = 3.44, p = 0.18$ ), suggesting that both achieve an acceptable level of accuracy for this task.

## 6.7 Discussions

### 6.7.1 Model Computational Performance: Human vs. Simulation Results

While neither model showed strong statistical significance in estimating the processing time, the CWA-informed model approached significance (see Figure 42), producing a broader range of processing times, with both higher and lower values, similar to the variability observed in human performance. Whereas the Basic Model shows a narrower distribution with limited variability in processing time

The broader distribution comes mainly from the Strategy Analysis embedded in the CWA-informed model's construction to capture human decision-making paths flexibly. The flexible paths with more or fewer rules affect the length of the alert processing time. For instance, following one of the streamlined strategy paths could lead to a decision using only 17 rules, whereas the pattern-based strategy could lead to a decision connecting over thirty rules. The different number of rules connected leads to variability in processing time. The application of ConTA's guidance on the rules' coordinating and switching mechanisms also contributes to improving the model's performance. Specifically, ConTA captures the flexibility of cognitive processes and thus informs the coordination of branched rules by different strategies. Together, the CWA-informed model better captures the variation in human decision-making paths than the Basic Model.

The Chi-square tests showed no significant differences between either model and the empirical data on triage decision patterns. This result suggests that both models achieved generally acceptable accuracy in reproducing participants' decision outcomes. However, it is not surprising, as

the task was mainly carried out within the defined pattern alert entries with clear criteria. Meanwhile, one notable discrepancy between model and human results lies in the imbalance between ‘log’ and ‘escalate’ decisions: the model tends to produce more ‘escalate’ decisions, whereas human participants more often choose ‘log’. This suggests that the model's rule construction may be somewhat more aggressive in triggering ‘escalate’ decisions than what is typically observed in human behavior. To address this imbalance, further refinement of the model's rule triggering ‘log’ and ‘escalate’ is needed. Another reason might be that most participants in Group 2 were novices, while the CWA-informed model incorporates insights from more experienced participants (Group 1). There may be some misalignment or unaccounted discrepancies between the two groups’ information processing and decision-making.

Beyond the improved performance resulting from the CWA-informed construction process, the outcomes of both models highlight the need for further refinement and raise additional concerns. First, the high variability in human processing time remains unclear whether that is due to external disturbances or extended thinking time. But one possibility is the different screen sizes used among participants (ranging from reported 13" to 27"), which could affect the visual distance required for search and shifting attention. In ACT-R, the default saccadic shift time is approximately 85ms per visual shift under standard conditions. Given the triaging task involving at most 20 visual shifts (representing the longest rule chains, see Table 18), larger screens can increase the physical and cognitive effort per shift, and cumulatively result in a 1-2 s deviation in total processing time with different screen sizes. This finding suggests that the specific experimental conditions, such as screen size, can significantly influence variability in ACT-R based modelling performance. In other words, high precision is required in the behavioral environment setup, especially in spatial details like the exact stimulus locations, to ensure effective modeling results. However, such spatial precision is not always achievable in dynamic or complex real-world conditions, which may limit the model’s ability to generalize human performance estimations across broader task scopes reliably.

Second, both models exhibit lower standard deviations in processing time compared to the data from human participants, indicating a more consistent pattern than that of humans. One explanation for this significant variance in human data is the complexity of the alert task itself. Although the task was decomposed into smaller modules (i.e., list-view defined-pattern alerts only, and per-alert processing time), each alert’s processing still involves a long sequence of actions and cognitive processing rules (with at least 27 rules interconnected, see Appendix B Table 14). For

human, the longer the decision-making path, the greater the accumulated variability and the more branched cognitive paths. For the modelling effort, this raises a critical concern about the model's ability to estimate human performance in complex, long-chain tasks accurately. Prior discussions on ACT-R's architectural constraints also recommend applying the model to small-scale tasks initially, where cognitive variability can be more effectively constrained and interpreted (N. Taatgen & Anderson, 2008).

### **6.7.2 Enhanced Model Construction and Application Through the Integration of CWA and ACT-R**

The CWA-informed construction process implemented our proposal of integrating CWA's analysis with the ACT-R model. This implementation effectively embeds the alert triage task environment and knowledge into the model's construction. The improved modeling process validates our conceptual analysis of the compatibility and enhancement of CWA and ACT-R, as discussed in Chapter 4.

This integration also addresses the challenges outlined in Chapter 3 concerning the direct application of ACT-R to cybersecurity contexts. By incorporating CWA, the model construction process is strengthened through the specific embedding of SOC task knowledge and a detailed task environment representation, supported by a more systematic domain analysis. The model thus achieves higher domain fidelity in the production-rule construction, and the information displayed on the interface is sufficient to simulate real-world task operations. Although this work limits the modeling scope to tier-1 analysts' operations, it still integrates key CWA-guided analyses. This integration already improves the model's ability to capture the flexibility and variability in real-world human analysis and behavior. Consequently, the model better captures human performance in typical cybersecurity defense tasks that require both the preliminary analysis of dense information and reaction-time actions.

On the other hand, cybersecurity applications commonly involve multiple stakeholders with dynamic interactions. The CWA analysis helps clarify the effective task scope across shifting perspectives within the complex system (e.g., backend supportive SOC analysts and in-vehicle human driver), enabling more accurate task specification for ACT-R simulations. CWA's system-level consideration could also support future extensions that the ACT-R model's dynamic simulation of interactions between multiple defensive roles and attackers. Such extensions work may further reveal system vulnerabilities and inform defensive design.

The enhancement does not critique the construction or outcomes of prior ACT-R models. Instead, it highlights the importance of striking a balance between task fidelity and generalizability when applying this solid cognitive architecture to domain-specific modeling. This trade-off is a recurring topic in computational cognitive modeling (Duch et al., 2008; Ritter et al., 2003): the main goal is to develop a stable and reusable task model, which conflicts with overly detailed, task-specific constructions that risk overfitting. Nonetheless, a useful model must still preserve the transferability of its predictions to real-world task contexts, as the ultimate aim of a predictive model is to estimate human performance in applied settings accurately.

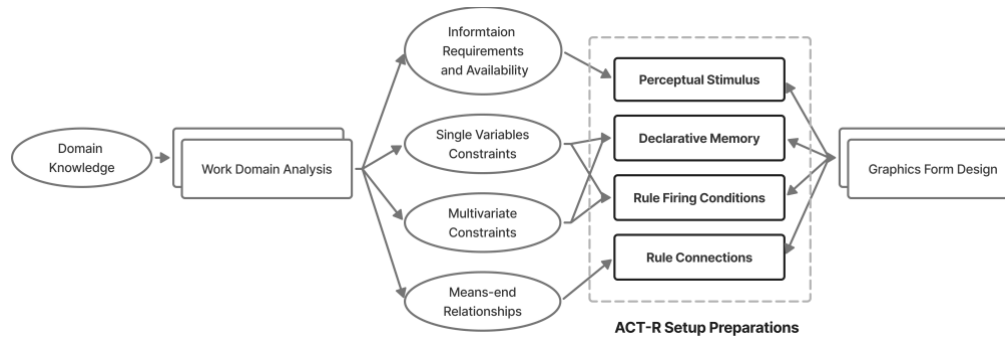
Beyond this study's integration of CWA's three dimensions to inform the ACT-R based model, there remains further potential to enrich task knowledge and environment. For instance, characteristics of the human operator, such as age and experience, are also essential aspects of the task environment. These factors could be further incorporated into the sub-symbolic mechanisms of ACT-R, enhancing the model's capacity to capture individual differences and improve its overall effectiveness. Future work could also explore how CWA insights may be used to refine these sub-symbolic components of the ACT-R model.

The benefit CWA gains from this modeling effort is a faster and more objective method of validating the qualitative insights, which is more prompt than relying only on domain experts' review and iterations. While expert validation is generally more reliable, it is usually time-consuming and poses challenges for those domain experts without prior CWA knowledge (Jenkins et al., 2007; Zhang & Lintern, 2024), making it difficult to follow a consistent validation framework. Computational modeling, by contrast, provides an opportunity to validate part of CWA's insights quickly. For example, in this study, the model's predicted decisions compared with humans' decisions help validate the comprehensiveness and tendencies identified in the CWA strategy analysis. Of course, some aspects of CWA results still cannot be validated computationally. Only a limited number of human behaviors are simulated by the model.

### **6.7.3 Interface Design and Cognitive Modelling**

In this experiment, we used a subset of WDA variables (see Section 6.5.4) as chunk slots in ACT-R's declarative memory. We referenced the relationships among these variables to construct the model's production rule conditions (see Figure 44). The specific values of these variables were presented as

visual elements on the user interface, allowing for a direct correspondence between the model's rules and the task environment.



**Figure 44.** From WDA to ACT-R Setup Preparations.

In modeling the sequence of processing these visual elements, we observed that human participants prioritize visually prominent elements over longer text-based information. Participants first directed their attention to key color-coded or numbered indicators. Notably, novice participants quickly develop a reliance on these perceptually prominent key indicators, forming a Streamlined Strategy for decision-making (e.g., always prioritizing severity levels and anomaly scores). This finding aligns with EID, suggesting that information displayed perceptually saliently for higher-level variables reduces cognitive load for prompt monitoring (C. M. Burns & Hajdukiewicz, 2017). The ACT-R models also explain this tendency through the rule-switch mechanism, which manages cognitive workload with an information loss tradeoff (C. Wu et al., 2008). When cognitive capability is low, human agents follow the most immediate and less demanding path, even if it is not necessarily the most accurate one.

This finding, supported by both cognitive modelling approaches, helps explain why novice participants tend to develop less reliable but lower-effort strategies, influenced by the design of the display interface. On the other hand, it also suggests that this overreliance tendency could be mitigated through counterbalancing design approaches that guide human decision-making back toward more stable, robust, and accurate strategies. This could be achieved by displaying the other alert features or indicators at the same or higher level of visual prominence, or by using more synthesized visualizations, such as plots or graphs, to enhance cognitive efficiency while avoiding overreliance on a single feature. However, the visualizations in our experimental interface were limited, as we did not incorporate formats that more effectively convey variable relationships, support

intuitive processing, or allow for comparison and historical tracking (C. M. Burns & Hajdukiewicz, 2017). This was partly influenced by ACT-R's focus on stimulus-driven inputs, which has challenges in modeling interactions with complex visualizations such as multi-dimensional graphs (Dimov et al., 2020). Still, the insights gained from tracking human performance through the cognitive modeling process represent a valuable contribution to human-centered design efforts.

#### **6.7.4 Modelling Scope: Modeling Knowledge-Based Decision-Making**

We did not model the Adaptive Strategy used in ambiguous alert decision-making. As illustrated in the Strategy Analysis (see Figure 33), the Adaptive Strategy has more branching and emerging decision pathways and a broader use of alert features. The challenge in modeling such knowledge-based (Rasmussen, 1985) decision-making lies in identifying diverse features and capturing the emerging exploratory paths. Rule-based models lack mechanisms for open-ended exploration, making it difficult to represent these dynamic paths exhaustively (Kotseruba & Tsotsos, 2020). This limitation reduces the model's exploratory power when handling ambiguous or unexpected conditions. While ACT-R includes a learning component, it remains fundamentally rule-based, meaning its reward mechanism optimizes behavior within predefined rule structures rather than evolving new solutions (N. A. Taatgen & Anderson, 2002). This constraint makes it challenging to quantitatively model flexible, knowledge-based decision-making in unanticipated scenarios, which is a common challenge for symbolic modeling approaches (Kotseruba & Tsotsos, 2020), as mentioned in Chapter 2, due to its lack of "creativity".

On the other hand, the rapid advancement of statistical pattern-based reasoning (e.g., neural networks and deep learning) has significantly enhanced the performance of AI models with greater exploratory and adaptive capabilities (Russell & Norvig, 2022). Given this shift in the cognitive modeling landscape, we conducted a follow-up study in Chapter 7 to investigate whether the integrated cognitive models can be enhanced by leveraging the AI models for predicting human performance in unanticipated scenarios.

#### **6.7.5 Aspects Not Captured by the Model**

Besides the mentioned limited modeling scope of human behaviors (Section 6.5.8), two notable human behaviors observed but not captured by the model are: (1) participants may change their

decisions after their initial attempt, and (2) participants considered factors beyond the predefined set of features.

Due to the challenges of statistically modelling the diversity of patterns in decision-making, our analysis focused only on participants' initial decisions and the corresponding reasoning and behaviors. However, we observed that quite a few participants (13 of the 21 who provided a complete list of valid responses) later revised their decisions after being exposed to more alert entries. In some cases, later alerts reminded them of previously encountered ones, prompting them to reevaluate earlier judgments. For instance, the overall distribution of their decisions appeared to influence subsequent choices. If participants noticed that they had escalated too many alerts, they sometimes reconsidered whether their prior criteria were overly strict or aggressive. The decision-changing pattern was not modeled in our model because the triggers for initiating the decision changes are implicit, and the time for rethinking and deciding the action changes is hard to capture behaviorally. Specifically, no consistent patterns were observed in terms of which factors led to participants' re-evaluation or re-action. A potential solution is to expand the interview for Group 1 participants to uncover patterns behind decision adjustments and actions.

Another interesting observation is that participants may consider factors beyond the information presented in the alert entries, particularly by expert participants. These additional considerations include contextual and environmental factors within their working environment. For example, *PII* shared that team workload allocation influences his/her triage decisions. Specifically, if Tier-2 analysts are short-staffed or overwhelmed, *PII* would be less likely to escalate an alert, unless it is very urgent. Such contextual factors may implicitly influence human decision-making strategies. Still, they were not explicitly defined in this study's work environment and were therefore excluded from the model's task environment setup. We did not include any factors beyond the information displayed on the interface in the model's task environment, which reflects a relatively narrow configuration for task environment analysis. This is also echoed by the analysis of this study's limitations in not covering all dimensions of CWA, which limits the comprehensiveness of the work domain analysis and the inclusion of important contextual factors enhancing task environment fidelity for model robustness.

## 6.8 Conclusions

Our motivation for integrating CWA and ACT-R was to develop an enhanced cognitive model that combines the complementary strengths of both for simulating human performance in complex work environments, such as the CAV cybersecurity domain, and to extend the cognitive model's development in supporting human-centered analysis and design across broader domains with socio-technical complexity.

Following the conceptual integration exploration in Chapter 4, this chapter presents the practical construction of the alert triage task cognitive model by systematically incorporating the formative analysis by CWA's three dimensions (WDA, ConTA, and Strategy Analysis) to the ACT-R based model construction (see Table 10). The modeling results yielded comparatively more accurate predictions of human task performance within the alert triage task. They identified potential flexibility and tendencies in human decision-making in a measurable and interpretable way. This confirms that the integration enhanced the model's construction efficiency and modelling performance.

**Table 10.** Integration of CWA and ACT-R with Enhancements.

Standard ACT-R	Informed by CWA	Enhanced Aspects
Declarative Modules	Work Domain Analysis	Formative Domain Analysis
Productions Rules Construction and Mapping	Control Task Analysis	Goal-Oriented Rule Connections
	Strategy Analysis	Flexibility & Interpretability of Rule Switching

Although the effort in this chapter is not a full implementation of all CWA dimensions integrated with the complete ACT-R architecture and mechanism, this effort represents the first ever attempt to combine these two classic cognitive modelling approaches in application.

## 6.9 Limitations and Future Works

Still, several compromises arise when attempting to align CWA and ACT-R for modelling.

The first limitation is in the task scope of the model's application. The shift from in-vehicle operation to a back-end external supportive sector limits the use of in-vehicle system information in the task and model construction. Given the final modeling's focus on only a subset of alerts, CAV-

related information was minimally incorporated. As a result, the experimental task model may align more closely with general SOC operations than with the specialized context of a VSOC.

Another limitation lies in the trade-off between implementing the model's quantitative capabilities and preserving the flexibility of human decision-making. Given the model's rule-based nature, we focused on simulating the direct processing of clear pattern alerts. Inevitably, the model simplifies the adaptability that humans demonstrate in handling ambiguous alerts and revising initial decisions. We did not fully explore ACT-R's sub-symbolic mechanisms, such as utility-based reinforcement learning, to better capture human adaptation and expertise development over time. The model also offers limited representation of visualization-based information processing, as its current mechanisms are not well-suited for this modality.

Third, as an initial exploration of model construction and application, the implementation may have omitted some details that could affect model precision. For instance, future studies should consider controlling the granularity of rule construction and the physical specifications of the task environment. Further refinement could involve decomposing subgoal rules to more precisely reflect the analytical process. There are demanding requirements for the task environment setup specifications to achieve precise control, as ACT-R is highly sensitive to spatial parameters for perceptual modules. Further refinement of the model's development and application should consider these factors to enhance its precision.

## 6.10 Summary

This chapter primarily aims to answer the fourth research question:

- *RQ3: What are the enhancements, limitations, and future directions of integrating CWA and ACT-R for modeling human performance in complex domains like cybersecurity?*

In short, this integration did enhance the model's construction efficiency, though the efficacy improvement is limited.

The integrated model demonstrated a more systematic and efficient model construction process for SOC analysts' alert triage operations. It also captured human analysts' decision-making flexibility and tendencies in a measurable and interpretable way. Overall, compared to without CWA, incorporating CWA insights into the ACT-R model enriches the task environment representation. It informs domain knowledge in the model construction, and thus improves the model's transferability,

and interpretability. On the other hand, the ACT-R model's results provide a direct, quantifiable and objective validation of CWA insights regarding how the task environment and domain knowledge shape human strategies. Therefore, the results confirm the fundamental compatibility of CWA and ACT-R and showcase the complementary benefits of their integration. In addition, this study details a systematic process for incorporating qualitative insights from CWA into a computational model. This model, therefore, increases domain specificity while preserving the generalizable computational capabilities.

In practical terms, this work extends cognitive modeling efforts to broader roles within the CAV domain, suggesting a defense framework that engages SOCs as external support. This extended consideration has the potential to inform the future development of more collaborative CAV cybersecurity strategies beyond in-vehicle operations. This shifted focus may further proactively facilitate the design of in-vehicle cybersecurity tools through interactions with external support systems.

However, as in this work, the computational model only transformed part of the CWA insights into building the model: there are still some limitations in fully capturing human performance. These limitations stem from trade-offs in balancing task complexity, modeling scope, and the fidelity and precision of the task environment. The challenge of translating exploratory knowledge-based decision-making into the generalized rule-based (if-then) construct persists.

To address this key challenge, a potential future direction is to incorporate Generative AI models (Malloy & Gonzalez, 2024), such as LLMs, as valuable complements to our cognitive model. In Appendix C, we therefore conducted a brief follow-up study using a representative GAI model to explore this potential. In the next concluding chapter, we will revisit the outcomes of the research questions, extend the conclusions from each chapter into an overarching discussion and the further advancement of cognitive modeling and applications.

## Chapter 7

### Summary and Future Works

This chapter will review the key findings, extend the discussion, acknowledge the study's limitations, present a brief follow-up exploring the potential of GAI models to address existing gaps, and conclude with suggested directions for future research.

The figure below recaps the thesis structure in Chapter 2.

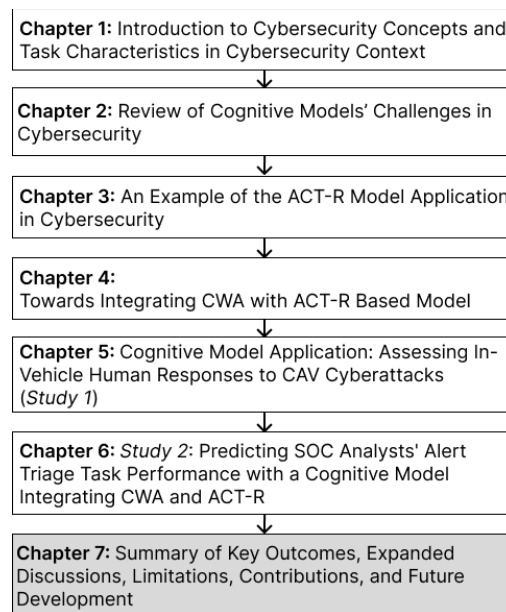


Figure 45. Overview of Thesis Structure – Chapter 7.

### 7.1 Summary of Key Outcomes

Cybersecurity is a trending concern with the rapid development of today's connected systems. While humans are often considered vulnerable targets, research on human factors remains limited compared to the extensive technical focus on defense and mitigation strategies. Human-focused cognitive research in this domain faces two main challenges: the dynamic and complex nature of the cybersecurity landscape and the domain-specific characteristics of systems that further shape human performance in responding to cyberattacks. These challenges point to the need for a cognitive analysis that can model and interpret human performance with both precise dynamic measurement and domain-specific fidelity.

Thus, to work toward such a model, we began by reviewing existing cognitive models and justified the selection of the ACT-R model due to its widespread application and validated capability in simulating human behaviors. To examine the model's direct applicability, we conducted a preliminary exploratory attempt using an extended ACT-R model alone to estimate in-vehicle human responses to distraction-based cyber threats, to answer our first research question:

***RQ1: Is an ACT-R model sufficient for modeling human performance in cybersecurity?***

Through the initial modelling attempt, we identified several key limitations in constructing and directly applying an ACT-R model to this domain: (1) the informal subjective process of model's rule construction without sufficient task-specific knowledge; (2) the simplified task environment with limited fidelity based on a generalized task model; and (3) the need of an effective task scope for cybersecurity management to demonstrate the model's strengths better.

In response, we proposed a solution by integrating CWA guided insights with an ACT-R model. We first introduced CWA's advantages in guiding formative analysis of domain-specific knowledge and increased task environment fidelity. A detailed elaboration on the CWA and ACT-R's structural compatibility across dimensions, their fundamental strengths as complements, and the functional competencies with integration was presented. This conceptual exploration demonstrated the feasibility of the CWA and ACT-R's integration and the improvement in the model construction efficiency and domain-specific validity to answer the second research question:

***RQ2: Can insights from Cognitive Work Analysis (CWA) enhance the effectiveness of an ACT-R cognitive model in complex domains such as cybersecurity?***

Building on the exploration of CWA and ACT-R's integration of domain-specific analysis with computational capabilities, we sought to apply the model in an effective cybersecurity management task environment to simulate human. We first considered to continue using in-vehicle operators' responses to cyberattacks as the model's application example. Hence, assessing in-vehicle human responses to cyber threats led to the sub-question of RQ2a.

***RQ2a. Is in-vehicle operation an effective task environment for detecting and mitigating cybersecurity threats?***

A survey of drivers' responses to different forms of hypothetical cybersecurity scenarios was conducted. The attack scenarios could impact the driving display, the motion of the vehicle or both.

We found, however, in general drivers demonstrated limited analytical processing when responding to perceived anomalies by cyber-attacks. Instead, drivers tended to rely on rapid and skill-based perceptual-motor decisions such as pulling over, restarting the system to sustain the driving safety, or seeking assistance from external authorities for direct guidance in handling cyber threats. These are all sensible responses, particularly given the minimal training, experience and information provided to vehicle drivers. Given the study results, in-vehicle operation alone could not be regarded as an effective environment for detecting or mitigating cybersecurity threats. We need a task environment where users were actually involved in threat detection and mitigation. Therefore, the focus was shifted to another cybersecurity application - the Security Operations Center (SOC) in coordinating defensive responses to support systems' end-users such as human drivers:

***RQ2b: Is operation within a Security Operations Center an effective task environment for detecting and mitigating cybersecurity threats?***

A Security Operations Center is widely recognized as a best practice for monitoring and coordinating cybersecurity defense strategies, and thus can be regarded as an effective task environment for cybersecurity management. Therefore, we constructed the integrated cognitive model by combining CWA-guided insights of SOC work domain with the ACT-R architecture to simulate its analysts' task performance in cybersecurity alert triage. The proposed model was developed from model construction preparation to detailed rule development, guided by CWA's Work Domain Analysis, Control Task Analysis, and Strategies Analysis. The model development process and simulation results with findings, therefore, provide the answer to the final research question:

***RQ3. What are the enhancements, limitations, and future directions of integrating CWA and ACT-R for modeling human performance in complex domains like cybersecurity?***

The integrated model showed more systematic construction process, its simulation results illustrated a trend toward improved quantitative accuracy and demonstrated enhanced domain-specific validity with greater interpretability of human adaptability and flexibility. These findings indicates that the integrated model could be a useful cognitive model for cybersecurity.

However, the model's construction has considered inevitable trade-offs in balancing computational generalization with human exploratory behavior and implicit strategy adjustments. As a result, the integrated model still faces limitations in fully capturing the breadth of human analyst performance across the entire task operations.

The following discussion will in detail discuss our efforts in developing the integrated model to address cognitive modeling challenges in cybersecurity, explaining how some challenges have been mitigated, summarizing the model's development from this work (see Section 7.3), and remaining gaps could be further explored in future studies (See Section 7.3.3).

## **7.2 Enhancements by CWA-Informed ACT-R Models in Cybersecurity**

The initial aim of this work is to explore a robust cognitive model in cybersecurity, beginning with an illustrative example of how in-vehicle human drivers respond to CAV cyberattacks. The review of cognitive model applications and the preliminary modeling attempt in driver responses to CAV distraction attacks revealed main challenges in constructing and applying current cognitive models within the cybersecurity domain: the need for domain-specific knowledge and sufficient fidelity of the cyberattack management task environment (Veksler et al., 2018). Modeling human decision-making in cybersecurity requires the cognitive model's improvements in both.

### **7.2.1 Cognitive Modeling Informed by Task Knowledge**

Cyberattacks differ from conventional malfunctions in many ways. Cyberattacks are usually stealthier than conventional faults, as they often involve intentional deception (Aliwa et al., 2020; M. Chen & Yan, 2023; Nowdehi et al., 2019). A broader view of the communication network and access to more data and information are critical for effectively responding to cyberattacks, whereas conventional malfunctions typically occur in isolation (Aliwa et al., 2020; Linkov et al., 2019). And, cyberattacks often affect multiple stakeholders and lead to cascading impacts given the highly interconnected nature (Kidmose, 2025; Vivek et al., 2019). These inherent differences can significantly affect human response effectiveness than conventional malfunctions, as delayed actions, overreactions, or passivity may lead to more risky consequences (M. Wang et al., 2024). This also results in a modeling difference between human handling cyberattacks and conventional malfunctions. Responding to cyberattacks requires specific cybersecurity knowledge that supports dynamic analytical information processing for detection and response. For instance, in Study 2, the runbook's basic criteria for cybersecurity alert triage contributed to the model's declarative knowledge and production rules supporting the model's plausibility in simulating analysts' decision-makings. Moreover, participants' cybersecurity knowledge influenced their analytic processing in alert feature selection, judgment in evaluating individual alert features, and assessment of alert feature combinations. Constructing the

model upon the cybersecurity knowledge to reflect its potential impact on decision-making enhances the model's validity and capacity to simulate variability in human performance.

On the other hand, since cybersecurity often functions as a supporting layer of complex socio-technical domains, human decision-making must also account for the domain's own goals and context. The specific domain knowledge is therefore also critical for modeling human decisions and behavioral patterns in managing cyber threats. For example, in study 2, some participants analyzed the vehicle system's motion status to determine whether an alert was critical beyond general alert feature patterns. The domain analysis of vehicle system knowledge can provide a deeper understanding of human decision-making regarding certain cybersecurity risks and contribute to refining the model for edge cases, such as when transportation safety risks are more severe than cybersecurity implications.

Consequently, integrating CWA's analysis of domain-specific knowledge into the cognitive model enhanced the model's plausibility in simulating human reasoning processes in response to cyberattacks.

### **7.2.2 Cognitive Modeling with Enhanced Task Environment Fidelity**

The advancement of the technological landscape in both cybersecurity and complex systems such as CAVs is increasingly characterized by intensive, information-driven processes (Biondi & Jajo, 2024; Muzahid et al., 2024). As such, overly generalized task environments will fail to capture the real-world complexity in modelling human performance in information processing, leading to oversimplification, reduced task fidelity, and limited transferability of the model's results.

In this work, task fidelity mainly refers to the domain-specific information in the alerts. The simulation of human decision-making is highly dependent on the analytical processing of detailed alert information provided to human operators. In Study 2, participants developed shortcuts and adaptive strategies based on the analysis of the criticality of the detailed alert features, such as severity levels and anomaly scores. With low task fidelity in information details, participants would not generate dynamic strategies in the alert triage; consequently, the dynamics of human performance could not be captured in the task modeling. In other words, the decision-making shortcuts and adaptive strategies could arise only from operators' interpretations of the specific alert information

that is supported by sufficient task fidelity. If only with abstract tasks representing generalized cognitive processes, such nuanced cognitive pathways are hard to capture in the model.

Another consideration of task fidelity is assessing how effectively the task environment supports human manage cyberthreats in reality. For example, in Study 1 we examined the task environment of in-vehicle operations, which proved to be an ineffective cybersecurity management context. Human operations are constrained by domain-specific factors such as limited communication and inadequate defensive support, making it difficult to establish an effective defense layer for CAV cybersecurity. So we turned our research focus to analysts' performance in SOCs, where task environment is more supportive of effective cybersecurity management. Such an assessment of the task environment's effectiveness in defending against cyber threats requires a practical examination with sufficient task fidelity to confirm the model's applicable focus.

At both ends, task fidelity was enhanced through CWA-guided domain-specific analysis, accounting for cybersecurity specific complexity and broader task environmental constraints, therefore improving the model's nuanced representation of human decision-making and adaptability for the dynamic task environment in cyberspace.

### **7.2.3 Cognitive Modeling with Further Development for Cybersecurity Applications**

The integrated model applied in this work remains limited to a subset of SOC tier-1 analyst operations. Although the task scope is limited, the alert triage task involves processes of analyzing presented information and interacting with an information-rich environment by performing operations under time pressure. These task characteristics mirror common features of the cybersecurity tasks.

As an initial attempt, the model's application is a simple interface-interaction task simulation that did not require extensive cybersecurity context beyond the predefined training instructions. This simple task, focused on individual analyst operations, serves as a starting point for introducing additional CWA dimensions and for the extended ACT-R models integrations, allowing future applications to leverage respective modeling strengths more fully. For instance, CWA is powerful in analyzing the interactions between roles as attackers and defenders to reveal the different intent assessments, which impact the exposure of information and operations. CWA-guided analysis is also effective in incorporating organizational considerations into the defense framework at a systemic level (C. M. Burns et al., 2005). Extended ACT-R models can build on this work as a foundation to

translate additional CWA system-level insights into computational measures, such as simulating defense team collaboration performance, modeling dynamic interactions among different roles, and capturing individual characteristics and task-environment constraints in cybersecurity defense and attacker interactions. These areas are, in fact, important topics in human-centric defense frameworks and cybersecurity strategy research (Grobler et al., 2021; Gutzwiller et al., 2020; Khadka & Ullah, 2025; Veksler et al., 2018).

The distinct strengths of the two approaches in addressing these core challenges form the primary motivation for the integration in this work for cybersecurity applications. This initial attempt serves only as a foundational step. The goal is to incorporate more dimensions from both CWA and ACT-R in future work to enhance their combined modeling performance for cognitive modeling in cybersecurity.

### **7.3 Development of Cognitive Models**

Beyond addressing challenges in cognitive modeling for cybersecurity, this work also contributes to the advancement of general cognitive models in similar, highly sociotechnical, and complex systems. This section highlights the key achievements and limitations of this work in the development of such a cognitive model.

#### **7.3.1 Integration of CWA and ACT-R**

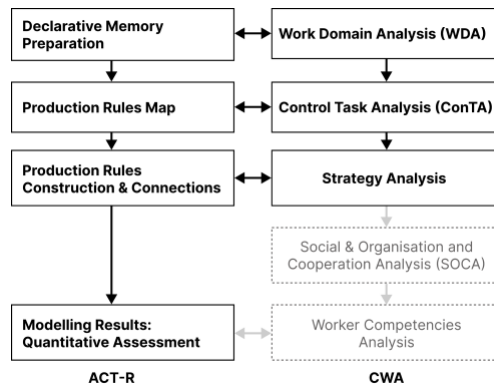
Study 2 elaborates that the three dimensions of CWA (WDA, ConTA, and Strategy Analysis) effectively support ACT-R model construction by systematically considering domain-specific constraints and enhancing the task fidelity of the ACT-R model. Beginning with WDA to identify information requirements and develop the SOC alert triage tool interface prototype, laying the foundation for the subsequent prototyping of the ConTA decision ladder and production rules coordination. The information requirements, constraints, and relationships derived from WDA serve as references for designing our SOC visual interfaces and preparing ACT-R modeling's perceptual elements and memory chunks (see Figure 44). The Strategy Analysis then in detail informs the rules connection and switching. The resulting model provides quantitative estimates of human performance in alert processing time and accuracy.

**Table 11.** An Integration of CWA and ACT-R with Achieved Enhancement.

ACT-R	CWA	Enhancement with CWA	Enhancement with ACT-R
Declarative Modules	Work Domain Analysis	Ecological Analysis of Work Environment	Quantifiable Validation of Domain-Specific Human Performance Analysis; Biologically Inspired Interpretation of Human Competence Constraints; Traceable temporal sequences of cognitive processes;
Productions Rule Map	Control Task Analysis	Information-to-Action & Action-to-Information Connections	
Productions Rule Construction and Connections	Strategy Analysis	Complexity and Flexibility of Decision-Making Paths	

Many of the human behaviors captured by the CWA, such as flexible decision-making pathways based on alert patterns, heuristic strategies using subsets of alert features, and reliance on perceptually salient interface cues, also informed the ACT-R model’s construction. These behaviors are reflected in the improved modeling results and can be represented through ACT-R’s core mechanisms, such as the sequence of visual search and the cognitive cost of memory retrieval processes underlying the strategies analysis. In this way, the integrated modelling not only provides measurable estimates but also offers traceable and interpretable evidence that supports CWA-guided contextual insights and identifies potential human vulnerability due to cognitive competence constraints.

Table 11 summarizes the proposed integrated modeling approaches, highlighting how the complementary strengths of CWA and ACT-R were combined. Figure 46 illustrates the integration of ACT-R model’s construction with a dimensional CWA framework (the dashed border boxes indicate phases not included in our work’s implementation).



**Figure 46.** Between CWA and ACT-R Model Construction.

### 7.3.2 Best Use Case of the Model

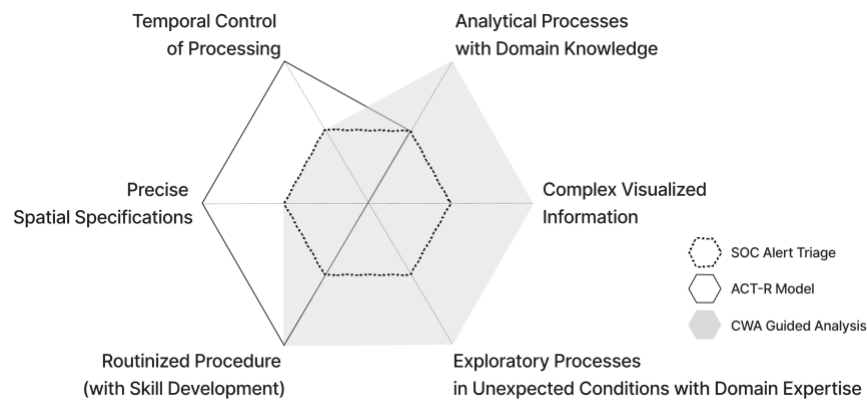
Although the integrated model demonstrates the above advancements in the construction process and results, we have also noted some constraints in its applicability.

Findings from Study 1 suggest that the in-vehicle environment may not be well-suited for leveraging the model's enhancements in predicting human responses to cyber threats, as analytic decision-making is largely absent in in-vehicle defensive responses. By comparison, a more suitable task environment for applying the proposed cognitive model should involve more analytical processes supported by domain knowledge, while still requiring time-sensitive responses, as illustrated in Study 2 on SOC analysts' tasks. While in study 2, we simulated SOC analysts' behaviors only in triage decisions upon clear pattern alerts within a specific interface design (i.e., list views of alert entries with defined patterns). The model did not expand to simulate more flexible reasoning processes and exploratory decision-making on ambiguous pattern alerts with additional information. The limited scope of the model's application primarily results from the distinctiveness between the ACT-R and CWA emphases in its application.

There are use cases where combining ACT-R and CWA does not provide a better model. In particular, some behaviors may be better modeled with ACT-R alone, while others may be better described with CWA. The ACT-R model is the most suitable model for behaviors involving temporal responses and well-defined perceptual-motor pathways (R. Xu et al., 2024) leading to final decisions. In comparison, analyzing human decision-making under unexpected conditions is the strength of CWA, but it is difficult to encode into the ACT-R model's rigid if-then rule structure. When handling ambiguous alert patterns, human participants exhibited large variability in their interpretation of alert

features, based on subjective experiences or expert knowledge, and developed dynamic decision-making strategies accordingly. These explorative and implicit adaptive decision-making may extend beyond the ACT-R model's most applicable task environment and knowledge base. Therefore, the integrated model, inheriting merely the overlapping applicability of both models, is not effective in handling all conditions.

Considering these limitations, we outline several task characteristics for the integrated model's best use cases. These characteristics are grounded in the areas where CWA and ACT-R are compatible but are also constrained by their differing emphases in analytical focus, as illustrated in Figure 47.



**Figure 47.** Task Environment for the Integrated Model's Applicability.

The ACT-R model is effective in estimating task processing time (i.e., Temporal Control of Processing) and requires relatively precise spatial specifications for accurate computation. It also excels in translating routine task procedures into symbolic model structures and modeling skill acquisition through its utility mechanisms. An ACT-R model can incorporate analytic processing in its production rule construction and allow some flexibility in adapting human analytic strategy selections, as shown in our Study 2. However, building production rules that represent the flexibility of human decision-making requires substantial analytical effort from the modeler. Especially when the analytic process of the task becomes highly complex or involves long chains of reasoning, fully represent the flexibility is not the ACT-R model's strength. These considerations influencing the applicable scope of the ACT-R model are illustrated in Figure 47, where the black solid-lined area represents the model's optimal performance. Notably, the basic ACT-R model is ineffective in

simulating explorative processes for unexpected situations (see Section 2.1.1) and has a limited ability to interpret complex visualized information (Nyamsuren & Taatgen, 2013).

Comparatively, CWA-guided analysis show strength in structuring domain-specific knowledge, including identifying specific task procedures, exploring in-depth analytical processing for decision-making, and supporting exploratory processes under unanticipated conditions. Based on its domain-specific analysis, CWA also contributes significantly to informing ecological interface design (see EID (C. M. Burns & Hajdukiewicz, 2017)) by visualizing complex information in ways that facilitate the human decision-making process. These strengths are illustrated on the right-hand side of Figure 35, represented by the grey-shaded area. Meanwhile, CWA focuses on work domain constraints, including temporal and spatial factors of the operational environment in some tasks. Therefore, the CWA's applicability extends into the left-hand side of the map, although the CWA-guided analysis of temporal and spatial constraints on human decisions differs from the specific requirements of the ACT-R model's application.

In modeling the alert triage task, part of the task scope falls within the overlapping strengths of CWA and ACT-R, as shown in the dashed-line area of Figure 35. This part of the task includes the runbook-guided standardized procedure following triage criteria on well-defined alert patterns, necessitating prompt responses, and with fixed spatial specifications within the display's list-view interface design. Whereas the other part of the triage task, where alerts do not follow predefined patterns, requires analysts to use domain expertise to handle ambiguous situations with diagnostic reasoning, and interpret complex visualizations of information beyond the list-view display. This part of the triage task falls outside the overlapping strengths of CWA and ACT-R, and therefore does not fit the best use case for the model. This explains why we use the model to capture analysts' decision-making only on well-defined patterned alerts.

It should also be noted that this does not imply the integrated model is inapplicable to tasks beyond the overlapped task features. For example, it could still be used to simulate in-vehicle drivers' reactions to cyber threats that do not involve analytical process; although in such cases, the model's performance would not differ a lot from that of a cognitive model without CWA insights. Likewise, the model could be applied to tasks without perceptual-motor procedures, such as post-hoc analysis of a cyber threat that does not require immediate action; however, in such tasks, the model would lose the advantage of ACT-R's computational accuracy in estimating action time.

In conclusion, when using the integrated model to simulate human performance, the main task characteristics should be evaluated to ensure for fully leveraging the model's enhancement.

### **7.3.3 Limitations and Future Development**

As this is an initial attempt to integrate the two cognitive modeling approaches, certain limitations remain. While individual limitations were discussed in earlier chapters, we have collected and summarized them here to provide a more comprehensive view:

#### **7.3.3.1 Limitations in Modeling Focus Shift from In-Vehicle to External Support Sector**

This modeling effort began by simulating the responses of in-vehicle human drivers to cyberattacks as an illustrative example for identifying cognitive modeling challenges. However, our survey-based investigation of drivers' responses to various forms of cyber threats indicates that in-vehicle decision-making tends to be less analytical, so in-vehicle responses was not an effective use case to pursue. Therefore, we then shifted the model application beyond in-vehicle operations to the remote SOC as an emerging constituent of defensive support. Tasks in this sector demand greater domain knowledge for nuanced decision-making to time-sensitive responses, making it a more effective task environment for demonstrating the enhanced model's strength in human performance simulation.

As the modeling focus shifted, we did not fully explore human vulnerabilities within the in-vehicle environment. Moreover, as a survey study with hypothetical threats, the results may differ from actual driver responses to real cyber threats during a tangible driving context. Future research could enhance the fidelity of the in-vehicle environment (e.g., by incorporating realistic attack scenarios with precise specifications) and increase domain specificity (e.g., through a detailed analysis of the representation of in-vehicle information clusters) to improve the validity of the analysis. Other research directions could focus on the cybersecurity situation awareness of different populations (e.g., age, driving and cybersecurity expertise, roles as passengers, and various vehicle types) under different driving conditions. Their varied responses to cyberattacks also could extend to research on developing customized communication-support tools for on-board cyber threat diagnostics, with guidance from external support sectors, to ensure appropriate responses for in-vehicle CSA improvement.

### 7.3.3.2 Limitations in Model Implementation Precision

As an initial exploration of an integrated model construction, the precision of the implementation could be further refined. In this work, the precision of rule construction was adjusted using a global parameter. Future development could improve this aspect by more precisely decomposing the current rules into finer subtasks for better precision alignment. Another limitation is the precision of the task environment's physical specification setup, as the ACT-R model is very demanding in terms of spatial parameters. The variability in screen size may introduce potential influences on temporal results, so future work should carefully consider the task environment's setup to enhance the model's precision, especially for modeling complex tasks with spatial demands.

### 7.3.3.3 Limitations in Modeling Human Affective States in Cybersecurity Contexts

Another limitation relates to molding human affective responses in cybersecurity contexts. Human end users, particularly those without sufficient background knowledge or operating under time pressure, often experience panic, stress, or other affective reactions that can impair judgment and increase risky behaviors (Nobles, 2022). The emotional and stress-induced responses remain a concern for understanding real-world human responses to cyber threats (Khadka & Ullah, 2025). However, ACT-R does not natively model panic, stress, or other affective conditions; its architecture is grounded in bounded-rational cognitive processes, and therefore cannot directly capture emotionally driven performance (Anderson, 2007).

However, the affective states can be partially captured through several CWA dimensions. Namely, the decision-making shortcuts embedded in the Decision Ladder, the rationale for strategy selection and switching in Strategy Analysis, the dynamics of team tension and trust in Social Organization and Cooperation, and individual traits and capabilities described in Worker Competencies. In addition, stress-related impacts could be introduced indirectly through time pressure to cognitive workload when multitasking demands arise during an attack. This points to a potential extension of the model example in Chapter 3, in which the ACT-R extended models could simulate competing attentional demands created by cyber disruptions leading to impacted human behaviors. Therefore, CWA-informed analysis can capture these affective factors and could in future work to extend native ACT-R extended models to incorporate affective influences.

#### 7.3.3.4 Limitations in Incomplete Use of ACT-R's mechanisms and CWA dimensions

This model did not leverage ACT-R's sub-symbolic mechanisms, such as chunk activation dynamics and utility-based reward mechanisms, to represent human factors like memory retrieval with different levels of expertise and the development of participant skills. For instance, more experienced participants would have higher base-level activation for relevant knowledge chunks, due to more frequent use or more recent exposure. Incorporating this consideration into the model may yield more nuanced results that vary with participants' levels of expertise. Also, the utility-based learning mechanism is currently used only for strategy selection and switching, but it could also be further adjusted to capture human adaptation and skill-based expertise development over time (e.g., production compilation) toward faster processing. However, these parameter adjustments and validations require further in-depth analysis with additional empirical data. As an initial modeling attempt focused on the integration attempt, this work does not yet include these more advanced considerations of the sub-symbolic mechanisms.

On the other hand, not all dimensions of CWA were implemented, potentially limiting the enhancement from the CWA insights. Specifically, this work only incorporates the first three dimensions of CWA (see Figure 46), the remaining two dimensions of SOCA (Social & Organizational and Cooperation Analysis), and Worker Competence Analysis are not included in this integration effort. The use of SOCA may support the development of a collaborative cybersecurity framework by guiding the analysis of different roles' responsibilities allocation, and communication coordination across end-users and supportive sectors. The Worker Competence Analysis could provide a systematic perspective for identifying gaps between the analysts' expected competence and simulated performance, helping to inform the design of facilitation tools and necessary training for the task. However, due to both dimensions limited use and lack of consistent application methods, they were not included in this integrated model at this stage.

Aside from the defensive focus, to more effectively capture the complexity of cyberspace dynamics in CAV cybersecurity, future cognitive model efforts should incorporate the interactions between both attacker and defenders' side, as emphasized in prior studies (Abbasi et al., 2015; Dominic et al., 2016; Hausken et al., 2024; Thackray et al., 2016; Veksler et al., 2018). This perspective is also one of the most nuanced compared to traditional cognitive modelling efforts in other domains, where adversarial interactions are minimal or absent. Although our work does not address this dimension in the modeling effort, CWA is particularly well-suited for analyzing how

multi-role entities with conflicting objectives and resource competition influence decision-making (C. M. Burns et al., 2005). Future cognitive modelling should incorporate this dimension of CWA more explicitly within cybersecurity contexts.

Beyond expanding the integration dimensions of both approaches to capture more complex behaviors and improving the model's implementation precision, a promising future development direction lies in incorporating fast-advancing GAI models to help bridge the gap in the model's exploratory capabilities. To explore this potential, we then conduct a brief comparison study to show the performance of a GAI model in predicting human decisions on ambiguous alert patterns (see Appendix C). The results will be compared with those of our classical modeling approach to discuss the future integration of GAI models into cognitive modeling development.

## **7.4 Guidance on the Application of the Integrated Model**

Section 7.3.2 outlines the model's best use case: it performs most effectively when cognitive processes and interaction patterns are clearly defined at the system level within a complex work environment, yet allow some flexibility (i.e., short paths may emerge through skill acquisition).

Building on this best use case, the following section presents a stepwise guide on how the integrated model could be better applied for human performance simulation from the experience of this work:

- 1. Identify an effective task environment. An effective task environment should first support human interaction to achieve the task objectives. A best effective task environment should also consider the key characteristics of the best-use-case scenarios discussed in Section 7.3.2.
- 2. Identify the task scope based on the WDA. Ensuring that the perceptual forms of tangible elements within the task environment are specified. Analysis based on domain work guidelines, regulatory and functional analysis documents, and SMEs consultations should inform the WDA. These validated WDA and the tangible elements within the task scope form the core of the ACT-R model's declarative chunks and the systematic representation of the task environment.

- 3. Confirm the specification of the task environment representation. This includes identifying precise spatial parameters, including the size, distance, and position of tangible operational elements. These specifications influence the precise simulation of processing time using Fitts' Law.
- 4. Construct the production rules through Strategy Analysis. Collecting operators interviews, behavioral tracking and interview for decision-making reasoning paths, translating the strategy analysis outcomes into production rule as condition-action pairs, incorporating the task environment elements identified through the WDA and ConTA.
- 5. Establish the task-specific workflow by the ConTA framework. The ConTA should be conducted by review of documents like task instructions and guidelines, prior validated experimental outcomes, and functional or operational manuals, and further supplemented and verified through insights from the WDA with system level analysis.
- 6. Examine the flexibility of decision-making strategies. Investigate the flexibility of decision-making paths through interviews with both novice and expert operators. Novices help reveal the flexibility and limitations of operational paths. Experts provide insight into the underlying reasoning, potential information processing short paths developed, and validation of the task environment and document-based ConTA. Together, these insights refine the ConTA results.
- 7. Map the Production Rules according to the ConTA. Production rules should be mapped and connected following the identified workflow and short-path variations, yielding a complete, interpretable, and traceable task model.
- 8. Run and refine the model through empirical validation. Collect data from a larger population to validate human-performance estimates and refine the model and use simulation traces to interpret human errors and limitations, informing iterative improvements from both an individual-capability and a broader system-design perspective.

This guidance summarizes the general advice for the CWA-informed ACT-R model construction and application by this work (see Table 12 for a brief guide to the methods and analyses for the model's respective phases), but the process does not require a strict sequential approach. For specific task simulation, the CWA framework and ACT-R model building procedure may be adapted,

as the CWA framework is formative and the ACT-R models are extensible. Their use can therefore be better aligned with the modeler’s judgment and the ongoing development of the two analysis methods.

**Table 12. Guidance on the Application of the Integrated Model.**

Step	Input	Output	Recommended Methods	Model Construction Phases
1	The task’s scope, objectives, and operational environment	Analysis of task objectives and the task characteristics alignment with the best use cases.	Modeler evaluation based on task analysis and field observations.	Identification of an effective task environment.
2	Work domain guidelines, regulatory and functional analysis documents.	The task scope and the within-scope tangible elements with their perceptual forms.	WDA based on domain work guidelines, regulatory and functional analysis documents, and SMEs consultations.	Preparation of declarative chunks and task environment elements.
3	The confirmed task environment and layout of operational elements	The precise spatial parameters: the element size, distance, and position	Quantification of the task environment’s spatial layout and operational element dimensions	Specifications of task environments, detailing the perceived elements dimensions and the operation’s starting position
4	Operational behaviors and the underlying decision-making reasoning	The condition–action pairs translated from the Strategy Analysis into production rules.	Strategy Analysis: transcribed and coded from operators' interviews about decision-making reasoning, behavioral tracking and observation	Construction of individual production rules as condition-action pairs over perceived elements or element combinations, resulting in actions or internal cognitive process updates.
5	The task-specific workflow derived from task instructions, documentation, and SMEs consultation	A ConTA linking identified strategies in a structured template of the cognitive flow for achieving the task goal	Analysis on task environment, operator experience, documents reviews, and transcribed SME interviews	ConTA mapped onto a Decision Ladder.
6	The divergent decision-making paths and operational deviations observed among operators.	Analysis of Flexible decision-making and the underlying reasoning for switching and selecting strategies based on conditions, individual	Analysis of interviews with operators examining the reasoning underlying their diverse cognitive pathways.	Refinement of the ConTA based on the identified divergent paths and the underlying reasoning.

		experience, and other contextual factors.		
7	The ConTA with cognitive path variations and underlying reasoning	A Production Rule Map incorporating diverse paths to achieve the task goal.	Mapping the strategy variations to the production rule map through adjustments to rule connections, condition specifications, and fallback rule conditions	The Production Rule Map for completing the task
8	Model simulation results and collected human empirical data	Comparison of the model's estimated outcomes with human performance	Data analysis	Model refinement through adjustments to rules definition, connections, parameter modifications, and the task environment

## 7.5 Contributions

This is known as the first attempt to integrate CWA and ACT-R in developing an enhanced cognitive model to simulate human performance, with a particular application in cybersecurity. This effort contributes to a more effective and efficient ACT-R model construction process with CWA's framework guided analysis, demonstrating the compatibility and the distinctiveness between the two classic modeling approaches. The process of translating CWA's domain-specific analysis into ACT-R model's core components elaborates a detailed pipeline for systematically incorporating domain-specific insights into a computational model. The resulting model also exhibits the trend towards improved predictive accuracy. Hence, this work advances the development of cognitive models with improved construction efficiency and predictive accuracy trend in complex domain as cybersecurity.

By examining the model's applicability, we identified how the two cognitive analysis approaches differ in their emphasis and application focuses. CWA is a formative approach that is more open to exploratory analytic processes with domain knowledge, where ACT-R provides the infrastructure for computational models to predict established human performance within cognitive capacities. The analysis of constraints by their integration informs the evaluation of the model's applicability, effectiveness, and task analysis dimensions across diverse tasks.

In practice, our cognitive modeling effort shifts focus from the isolated end users (e.g., in-vehicle drivers) to the broader collaborative cybersecurity defense framework's supportive roles (i.e., analysts in SOCs). The extended modeling focus potentially contributes to the advancement of

collaborative cybersecurity defense systems, including the design and development of the coordination and communication between centralized monitoring centers and end-users.

Lastly, this study evaluated the potential of utilizing an LLM to complement the integrated classic cognitive models' exploratory capability through a follow-up study. This comparison study revealed the performance deviations and respective strengths between GAI models versus traditional cognitive models. The discussion of using GAI models to simulate human decision-making broadens the perspective on the future development of cognitive models. By drawing on the insights from classic cognitive models, it also contributes to advancing AI models by offering guidance on improving error tracing, adaptive learning, feature selection, and prompt design optimization in broader applications.

## **7.6 Conclusions**

This thesis represents the first exploration and implementation of integrating CWA-guided domain-specific analysis with ACT-R's computational capabilities to develop an integrated cognitive model in complex work domains, with a particular application in cybersecurity domain. The integrated model was constructed and applied to predict the SOC analysts' performance in cyber alert triage tasks, demonstrating enhanced systematic efficiency in model construction and a trend toward improved simulation accuracy with domain-specificity. The work thus advances cognitive modeling by illustrating the combination of domain-specific analysis with precise computational cognitive models from theoretical grounding to practical implementations. The development of this cognitive model also sheds light on how classic models can inform the integration and improvement of GAI models in cognitive modeling.

## References

- Abbasi, Y. D., Short, M., Sinha, A., Sintov, N., Zhang, C., & Tambe, M. (2015). *Human Adversaries in Opportunistic Crime Security Games: Evaluating Competing Bounded Rationality Models*.
- Abosata, N., Al-Rubaye, S., Inalhan, G., & Emmanouilidis, C. (2021). Internet of Things for System Integrity: A Comprehensive Survey on Security, Attacks and Countermeasures for Industrial Applications. *Sensors*, 21(11), 3654. <https://doi.org/10.3390/s21113654>
- Abrar, M. M., Youssef, A., Islam, R., Satam, S., Latibari, B. S., Hariri, S., Shao, S., Salehi, S., & Satam, P. (2024). *GPS-IDS: An Anomaly-based GPS Spoofing Attack Detection Framework for Autonomous Vehicles* (arXiv:2405.08359). arXiv. <https://doi.org/10.48550/arXiv.2405.08359>
- Adams, S. S., Arel, I., Bach, J., Coop, R., Furlan, R., Goertzel, B., Hall, J. S., Samsonovich, A., Scheutz, M., Schlesinger, M., Shapiro, S. C., & Sowa, J. F. (2012). Mapping the Landscape of Human-Level Artificial General Intelligence. *AI Magazine*, 33(1), 25–41. <https://doi.org/10.1609/aimag.v33i1.2322>
- Agashe, S., Han, J., Gan, S., Yang, J., Li, A., & Wang, X. E. (2024). *Agent S: An Open Agentic Framework that Uses Computers Like a Human* (arXiv:2410.08164). arXiv. <https://doi.org/10.48550/arXiv.2410.08164>
- Aggarwal, P., Moisan, F., Gonzalez, C., & Dutt, V. (2022). Learning About the Effects of Alert Uncertainty in Attack and Defend Decisions via Cognitive Modeling. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 64(2), 343–358. <https://doi.org/10.1177/0018720820945425>
- Agyepong, E., Cherdantseva, Y., Reinecke, P., & Burnap, P. (2020). Towards a Framework for Measuring the Performance of a Security Operations Center Analyst. *2020 International*

- Conference on Cyber Security and Protection of Digital Services (Cyber Security)*, 1–8.  
<https://doi.org/10.1109/CyberSecurity49315.2020.9138872>
- Alahmadi, B. A., Axon, L., & Martinovic, I. (2022). *99% False Positives: A Qualitative Study of SOC Analysts' Perspectives on Security Alarms*.
- Alawadhi, M., Almazrouie, J., Kamil, M., & Khalil, K. A. (2020). Review and analysis of the importance of autonomous vehicles liability: A systematic literature review. *International Journal of System Assurance Engineering and Management*, *11*(6), 1227–1249.  
<https://doi.org/10.1007/s13198-020-00978-9>
- Aliwa, E., Rana, O., Perera, C., & Burnap, P. (2020). *Cyberattacks and Countermeasures For In-Vehicle Networks* (arXiv:2004.10781). arXiv. <https://doi.org/10.48550/arXiv.2004.10781>
- Alsaade, F. W., & Al-Adhaileh, M. H. (2023). Cyber Attack Detection for Self-Driving Vehicle Networks Using Deep Autoencoder Algorithms. *Sensors*, *23*(8), 4086.  
<https://doi.org/10.3390/s23084086>
- Alsharif, M., Mishra, S., & AlShehri, M. (2022). Impact of Human Vulnerabilities on Cybersecurity. *Computer Systems Science and Engineering*, *40*(3), 1153–1166.  
<https://doi.org/10.32604/csse.2022.019938>
- Alves, F., Bettini, A., Ferreira, P. M., & Bessani, A. (2021). Processing tweets for cybersecurity threat awareness. *Information Systems*, *95*, 101586. <https://doi.org/10.1016/j.is.2020.101586>
- Amabile, T. M. (1983). *The Social Psychology of Creativity*. Springer New York.  
<https://doi.org/10.1007/978-1-4612-5533-8>
- Anderson, J. R. (2007). *How Can the Human Mind Occur in the Physical Universe?* Oxford University Press. <https://doi.org/10.1093/acprof:oso/9780195324259.001.0001>

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036–1060. <https://doi.org/10.1037/0033-295X.111.4.1036>
- Anderson, J. R., Carter, C. S., Fincham, J. M., Qin, Y., Ravizza, S. M., & Rosenberg-Lee, M. (2008). Using fMRI to Test Models of Complex Cognition. *Cognitive Science*, *32*(8), 1323–1348. <https://doi.org/10.1080/03640210802451588>
- Anderson, J. R., & Schunn, C. D. (2000). *Implications of the ACT-R Learning Theory: No Magic Bullets*.
- Andrade, R. O., & Yoo, S. G. (2019). Cognitive security: A comprehensive study of cognitive science in cybersecurity. *Journal of Information Security and Applications*, *48*, 102352. <https://doi.org/10.1016/j.jisa.2019.06.008>
- Andreassen, J., Eileraas, M., Herrera, L. C., & Noori, N. S. (2023). InCREASE: A Dynamic Framework Towards Enhancing Situational Awareness in Cyber Incident Response. In T. Gjørseter, J. Radianti, & Y. Murayama (Eds.), *Information Technology in Disaster Risk Reduction* (Vol. 672, pp. 230–243). Springer Nature Switzerland. [https://doi.org/10.1007/978-3-031-34207-3\\_15](https://doi.org/10.1007/978-3-031-34207-3_15)
- Applebaum, A., Johnson, S., Limiero, M., & Smith, M. (2018). Playbook Oriented Cyber Response. *2018 National Cyber Summit (NCS)*, 8–15. <https://doi.org/10.1109/NCS.2018.00007>
- Awad, E., Levine, S., Kleiman-Weiner, M., Dsouza, S., Tenenbaum, J. B., Shariff, A., Bonnefon, J.-F., & Rahwan, I. (2018). *Blaming humans in autonomous vehicle accidents: Shared responsibility across levels of automation* (arXiv:1803.07170). arXiv. <https://doi.org/10.48550/arXiv.1803.07170>

- Ayodeji, A., Mohamed, M., Li, L., Di Buono, A., Pierce, I., & Ahmed, H. (2023). Cyber security in the nuclear industry: A closer look at digital control systems, networks and human factors. *Progress in Nuclear Energy*, *161*, 104738. <https://doi.org/10.1016/j.pnucene.2023.104738>
- Bachute, M. R., & Subhedar, J. M. (2021). Autonomous Driving Architectures: Insights of Machine Learning and Deep Learning Algorithms. *Machine Learning with Applications*, *6*, 100164. <https://doi.org/10.1016/j.mlwa.2021.100164>
- Backlinko Team. (2025). *Tesla Sales, Revenue & Production Statistics*. <https://backlinko.com/tesla-stats>
- Ban, G., & Jeon, M. (2025). Investigating drivers' responses to cyber-attacks while conducting non-driving related tasks in highly automated vehicles. *International Journal of Human-Computer Studies*, *202*, 103554. <https://doi.org/10.1016/j.ijhcs.2025.103554>
- Barletta, V. S., De Vincentiis, M., Buono, P., Pagano, A., & Caivano, D. (2023). Detecting attacks on in-vehicle networks through a mobile app. *2023 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT)*, 148–153. <https://doi.org/10.1109/3ICT60104.2023.10391747>
- Beckers, N., Siebert, L. C., Bruijnes, M., Jonker, C., & Abbink, D. (2022). Drivers of partially automated vehicles are blamed for crashes that they cannot reasonably avoid. *Scientific Reports*, *12*(1), 16193. <https://doi.org/10.1038/s41598-022-19876-0>
- Bilal, A., Ebert, D., & Lin, B. (2025). *LLMs for Explainable AI: A Comprehensive Survey* (arXiv:2504.00125). arXiv. <https://doi.org/10.48550/arXiv.2504.00125>
- Biondi, F. N., & Jajo, N. (2024). On the impact of on-road partially-automated driving on drivers' cognitive workload and attention allocation. *Accident Analysis & Prevention*, *200*, 107537. <https://doi.org/10.1016/j.aap.2024.107537>

- Birrell, S. A., Young, M. S., Jenkins, D. P., & Stanton, N. A. (2012). Cognitive Work Analysis for safe and efficient driving. *Theoretical Issues in Ergonomics Science*, *13*(4), 430–449.  
<https://doi.org/10.1080/1463922X.2010.539285>
- Bothell, D. (2004). *ACT-R 6.0 Reference Manual*.
- Bothell, D. (2022). *Act-r 7: Tutorial units-unit 7*. [act-r.psy.cmu.edu](http://act-r.psy.cmu.edu)
- Bridges, R. A., Rice, A. E., Oesch, S., Nichols, J. A., Watson, C., Spakes, K., Norem, S., Huettel, M., Jewell, B., Weber, B., Gannon, C., Bizovi, O., Hollifield, S. C., & Erwin, S. (2023). Testing SOAR tools in use. *Computers & Security*, *129*, 103201.  
<https://doi.org/10.1016/j.cose.2023.103201>
- Burns, C. (2013). Cognitive Work Analysis: New Dimensions. In P. Campos, T. Clemmensen, J. A. Nocera, D. Katre, A. Lopes, & R. Ørngreen (Eds.), *Human Work Interaction Design. Work Analysis and HCI* (Vol. 407, pp. 1–11). Springer Berlin Heidelberg.  
[https://doi.org/10.1007/978-3-642-41145-8\\_1](https://doi.org/10.1007/978-3-642-41145-8_1)
- Burns, C., Enomoto, Y., & Momtahan, K. (2008). A Cognitive Work Analysis of Cardiac Care Nurses Performing Teletriage. In A. Bisantz & C. Burns (Eds.), *Applications of Cognitive Work Analysis* (pp. 149–174). CRC Press. <https://doi.org/10.1201/9781420063059.ch7>
- Burns, C. M., Bryant, D. J., & Chalmers, B. A. (2005). Boundary, Purpose, and Values in Work-Domain Models: Models of Naval Command and Control. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, *35*(5), 603–616.  
<https://doi.org/10.1109/TSMCA.2005.851153>
- Burns, C. M., & Hajdukiewicz, J. (2017). *Ecological Interface Design* (1st ed.). CRC Press.  
<https://doi.org/10.1201/9781315272665>
- Byrne, M. D. (2012). *Cognitive Architectures in HCI: Present Work and Future Directions*.

- Byrne, M. D., & Gray, W. D. (2003). Returning Human Factors to an Engineering Discipline: Expanding the Science Base through a New Generation of Quantitative Methods - Preface to the Special Section. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(1), 1–4. <https://doi.org/10.1518/hfes.45.1.1.27229>
- Cao, S., & Liu, Y. (2013a). Queueing Network-ACTR Modeling of Concurrent Tasks Involving Multiple Controlled Processes. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 768–772. <https://doi.org/10.1177/1541931213571168>
- Cao, S., & Liu, Y. (2013b). Queueing network-adaptive control of thought rational (QN-ACTR): An integrated cognitive architecture for modelling complex cognitive and multi-task performance. *International Journal of Human Factors Modelling and Simulation*, 4(1), 63. <https://doi.org/10.1504/IJHFMS.2013.055790>
- Cao, S., & Liu, Y. (2014). Modeling Driving and Sentence Comprehension Dual-task Performance in Queueing Network-ACTR. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 808–811. <https://doi.org/10.1177/1541931214581170>
- Cao, S., & Liu, Y. (2015). Modelling workload in cognitive and concurrent tasks with time stress using an integrated cognitive architecture. *International Journal of Human Factors Modelling and Simulation*, 5(2), 113. <https://doi.org/10.1504/IJHFMS.2015.075360>
- Cao, S., Qin, Y., Jin, X., Zhao, L., & Shen, M. (2014). Effect of driving experience on collision avoidance braking: An experimental investigation and computational modelling. *Behaviour & Information Technology*, 33(9), 929–940. <https://doi.org/10.1080/0144929X.2014.902100>
- Cao, S., Qin, Y., Zhao, L., & Shen, M. (2015). Modeling the development of vehicle lateral control skills in a cognitive architecture. *Transportation Research Part F: Traffic Psychology and Behaviour*, 32, 1–10. <https://doi.org/10.1016/j.trf.2015.04.010>

- Champion, M., Jariwala, S., Ward, P., & Cooke, N. J. (2014). Using Cognitive Task Analysis to Investigate the Contribution of Informal Education to Developing Cyber Security Expertise. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 58(1), 310–314. <https://doi.org/10.1177/1541931214581064>
- Chen, M., & Yan, M. (2023). How to protect smart and autonomous vehicles from stealth viruses and worms. *ISA Transactions*, 141, 52–58. <https://doi.org/10.1016/j.isatra.2023.04.019>
- Chen, X., Wang, L., You, M., Liu, W., Fu, Y., Xu, J., Zhang, S., Chen, G., Li, K., & Li, J. (2024). *Evaluating and Enhancing Large Language Models' Performance in Domain-Specific Medicine: Development and Usability Study With DocOA (Preprint)*. <https://doi.org/10.2196/preprints.58158>
- Chong, H.-Q., Tan, A.-H., & Ng, G.-W. (2007). Integrated cognitive architectures: A survey. *Artificial Intelligence Review*, 28(2), 103–130. <https://doi.org/10.1007/s10462-009-9094-9>
- Christoforou, Z., Karlaftis, M. G., & Yannis, G. (2013). Reaction times of young alcohol-impaired drivers. *Accident Analysis & Prevention*, 61, 54–62. <https://doi.org/10.1016/j.aap.2012.12.030>
- Colabianchi, S., Costantino, F., Nonino, F., & Palombi, G. (2025). Transforming threats into opportunities: The role of human factors in enhancing cybersecurity. *Journal of Innovation & Knowledge*, 10(3), 100695. <https://doi.org/10.1016/j.jik.2025.100695>
- Collard, G., Ducroquet, S., Disson, E., & Talens, G. (2017). A definition of Information Security Classification in cybersecurity context. *2017 11th International Conference on Research Challenges in Information Science (RCIS)*, 77–82. <https://doi.org/10.1109/RCIS.2017.7956520>
- Collins, K. M., Wong, C., Feng, J., Wei, M., & Tenenbaum, J. B. (2022). *Structured, flexible, and robust: Benchmarking and improving large language models towards more human-like*

- behavior in out-of-distribution reasoning tasks* (arXiv:2205.05718). arXiv.  
<https://doi.org/10.48550/arXiv.2205.05718>
- Cornelissen, M., Salmon, P. M., McClure, R., & Stanton, N. A. (2013). Using cognitive work analysis and the strategies analysis diagram to understand variability in road user behaviour at intersections. *Ergonomics*, *56*(5), 764–780. <https://doi.org/10.1080/00140139.2013.768707>
- Coventry, L., & Branley, D. (2018). Cybersecurity in healthcare: A narrative review of trends, threats and ways forward. *Maturitas*, *113*, 48–52. <https://doi.org/10.1016/j.maturitas.2018.04.008>
- D’Amico, A., Whitley, K., Tesone, D., & O’Brien, B. (2005). Achieving Cyber Defense Situational Awareness: A Cognitive Task Analysis of Information Assurance Analysts. *The ANNUAL MEETING*.
- Deng, C., Cao, S., Wu, C., & Lyu, N. (2019a). Modeling Driver Take-Over Reaction Time and Emergency Response Time using an Integrated Cognitive Architecture. *Transportation Research Record: Journal of the Transportation Research Board*, *2673*(12), 380–390.  
<https://doi.org/10.1177/0361198119842114>
- Deng, C., Cao, S., Wu, C., & Lyu, N. (2019b). Predicting drivers’ direction sign reading reaction time using an integrated cognitive architecture. *IET Intelligent Transport Systems*, *13*(4), 622–627.  
<https://doi.org/10.1049/iet-its.2018.5160>
- Deng, C., Wu, C., Cao, S., & Lyu, N. (2019). Modeling the effect of limited sight distance through fog on car-following performance using QN-ACTR cognitive architecture. *Transportation Research Part F: Traffic Psychology and Behaviour*, *65*, 643–654.  
<https://doi.org/10.1016/j.trf.2017.12.017>
- Dimov, C., Khader, P. H., Marewski, J. N., & Pachur, T. (2020). How to model the neurocognitive dynamics of decision making: A methodological primer with ACT-R. *Behavior Research Methods*, *52*(2), 857–880. <https://doi.org/10.3758/s13428-019-01286-2>

- Ding, Y., Ying, J., Feng, Y., & Du, N. (2025). Explanations Help: Leveraging Human Capabilities to Detect Cyberattacks on Automated Vehicles. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–15. <https://doi.org/10.1145/3706598.3714301>
- Dominic, D., Chhawri, S., Eustice, R. M., Ma, D., & Weimerskirch, A. (2016). Risk Assessment for Cooperative Automated Driving. *Proceedings of the 2nd ACM Workshop on Cyber-Physical Systems Security and Privacy*, 47–58. <https://doi.org/10.1145/2994487.2994499>
- Duch, W., Oentaryo, R. J., & Pasquier, M. (2008). *Cognitive Architectures: Where do we go from here?* 171, 122–136.
- Dudziak, A., Kuranc, A., Zając, G., Szyszlak-Bargłowicz, J., Słowik, T., Stopka, O., Drożdziel, P., & Stopková, M. (2024). Smart solutions in car dashboard interfaces as a response to needs of drivers and their assessment. *Case Studies on Transport Policy*, 16, 101194. <https://doi.org/10.1016/j.cstp.2024.101194>
- Dukes, C. W. (2016). *Committee on National Security Systems(CNSS) GlossaryTHIS DOCUMENT PRESCRIBES MINIMUM STANDARDSYOUR DEPARTMENT OR AGENCY MAY REQUIRE FURTHERIMPLEMENTATION.*
- Durlik, I., Miller, T., Kostecka, E., Zwierzewicz, Z., & Łobodzińska, A. (2024). Cybersecurity in Autonomous Vehicles—Are We Ready for the Challenge? *Electronics*, 13(13), 2654. <https://doi.org/10.3390/electronics13132654>
- Dutt, V., Ahn, Y.-S., & Gonzalez, C. (2013). Cyber Situation Awareness: Modeling Detection of Cyber Attacks With Instance-Based Learning Theory. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 55(3), 605–618. <https://doi.org/10.1177/0018720812464045>

- Dutt, V., Moisan, F., & Gonzalez, C. (2016). Role of Intrusion-Detection Systems in Cyber-Attack Detection. In D. Nicholson (Ed.), *Advances in Human Factors in Cybersecurity* (Vol. 501, pp. 97–109). Springer International Publishing. [https://doi.org/10.1007/978-3-319-41932-9\\_9](https://doi.org/10.1007/978-3-319-41932-9_9)
- Eldardiry, O. M., & Caldwell, B. S. (2015). Improving Information and Task Coordination in Cyber Security Operation Centers. *IIE Annual Conference*, 1224–1233.  
<https://www.proquest.com/scholarly-journals/improving-information-task-coordination-cyber/docview/1791990375/se-2>
- Elliott, D., Keen, W., & Miao, L. (2019). Recent advances in connected and automated vehicles. *Journal of Traffic and Transportation Engineering (English Edition)*, 6(2), 109–131.  
<https://doi.org/10.1016/j.jtte.2018.09.005>
- Ellis, T., Balenson, D., & Locasto, M. (2022). Cyber Security Awareness Requirements for Operational Technology Systems. In J. Staggs & S. Sheno (Eds.), *Critical Infrastructure Protection XV* (Vol. 636, pp. 23–44). Springer International Publishing.  
[https://doi.org/10.1007/978-3-030-93511-5\\_2](https://doi.org/10.1007/978-3-030-93511-5_2)
- Federal Highway Administration. (2018). *CONNECTED AND AUTOMATED VEHICLES*. U.S. Department of Transportation.  
[https://www.planning.dot.gov/planning/topic\\_CVAV.aspx?utm\\_source](https://www.planning.dot.gov/planning/topic_CVAV.aspx?utm_source)
- Ferguson-Walter, K. J., Major, M. M., Johnson, C. K., Johnson, C. J., Scott, D. D., Gutzwiller, R. S., & Shade, T. (2023). Cyber expert feedback: Experiences, expectations, and opinions about cyber deception. *Computers & Security*, 130, 103268.  
<https://doi.org/10.1016/j.cose.2023.103268>
- Fidel, R., & Pejtersen, A. M. (2004). *From information behaviour research to the design of information systems: The Cognitive Work Analysis framework*.

- Flower, L., & Hayes, J. R. (1981). A Cognitive Process Theory of Writing. *College Composition and Communication*.
- Forsberg, J. (2022). *MEASURING THE TECHNICAL PERFORMANCE OF A SECURITY OPERATIONS CENTER*.
- Gamilla, A. P., & Palaoag, T. D. (2022). Building A Barrier: A Security Operations Center Framework For A Sustainable Smart Campus Network. *2022 6th International Conference on Information Technology (InCIT)*, 256–261.  
<https://doi.org/10.1109/InCIT56086.2022.10067377>
- Garza-Ulloa, J. (2022). Biomedical engineering and the evolution of artificial intelligence. In *Applied Biomedical Engineering Using Artificial Intelligence and Cognitive Models* (pp. 1–37). Elsevier. <https://doi.org/10.1016/B978-0-12-820718-5.00009-X>
- Gersh, J. R., McKneely, J. A., & Remington, R. W. (2005). Cognitive Engineering: Understanding Human Interaction with Complex Systems. *Johns Hopkins APL Technical Digest*, 26(4).
- Gershon, P., Sita, K. R., Zhu, C., Ehsani, J. P., Klauer, S. G., Dingus, T. A., & Simons-Morton, B. G. (2019). Distracted Driving, Visual Inattention, and Crash Risk Among Teenage Drivers. *American Journal of Preventive Medicine*, 56(4), 494–500.  
<https://doi.org/10.1016/j.amepre.2018.11.024>
- Gillmore, S. C., & Tenhundfeld, N. L. (2020). The Good, The Bad, and The Ugly: Evaluating Tesla’s Human Factors in the Wild West of Self-Driving Cars. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 64(1), 67–71.  
<https://doi.org/10.1177/1071181320641020>
- Gold, C., Damböck, D., Lorenz, L., & Bengler, K. (2013). “Take over!” How long does it take to get the driver back into the loop? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 57(1), 1938–1942. <https://doi.org/10.1177/1541931213571433>

- Gonzalez, C., Dutt, V., & Lebiere, C. (2013). Validating instance-based learning mechanisms outside of ACT-R. *Journal of Computational Science*, 4(4), 262–268.  
<https://doi.org/10.1016/j.jocs.2011.12.001>
- Grobler, M., Gaire, R., & Nepal, S. (2021). User, Usage and Usability: Redefining Human Centric Cyber Security. *Frontiers in Big Data*, 4, 583723. <https://doi.org/10.3389/fdata.2021.583723>
- Gupta, S., Maple, C., & Passerone, R. (2023). An Investigation of Cyber-Attacks and Security Mechanisms for Connected and Autonomous Vehicles. *IEEE Access*, 11, 90641–90669.  
<https://doi.org/10.1109/ACCESS.2023.3307473>
- Gutzwiller, R., Dykstra, J., & Payne, B. (2020). Gaps and Opportunities in Situational Awareness for Cybersecurity. *Digital Threats: Research and Practice*, 1(3), 1–6.  
<https://doi.org/10.1145/3384471>
- Hadlington, L., & Murphy, K. (2018). Is Media Multitasking Good for Cybersecurity? Exploring the Relationship Between Media Multitasking and Everyday Cognitive Failures on Self-Reported Risky Cybersecurity Behaviors. *Cyberpsychology, Behavior, and Social Networking*, 21(3), 168–172. <https://doi.org/10.1089/cyber.2017.0524>
- Hamad, M., Finkenzeller, A., Kühr, M., Roberts, A., Maennel, O., Prevelakis, V., & Steinhorst, S. (2024). REACT: Autonomous Intrusion Response System for Intelligent Vehicles. *Computers & Security*, 145, 104008. <https://doi.org/10.1016/j.cose.2024.104008>
- Han, F., Guo, L., Cui, H., & Lyu, Z. (2025). *Question Tokens Deserve More Attention: Enhancing Large Language Models without Training through Step-by-Step Reading and Question Attention Recalibration* (arXiv:2504.09402). arXiv.  
<https://doi.org/10.48550/arXiv.2504.09402>

- Han, J., Ju, Z., Chen, X., Yang, M., Zhang, H., & Huai, R. (2023). Secure Operations of Connected and Autonomous Vehicles. *IEEE Transactions on Intelligent Vehicles*, 8(11), 4484–4497. <https://doi.org/10.1109/TIV.2023.3304762>
- Harry, C., & Gallagher, N. (2025). *Classifying Cyber Events*.
- Hassall, M. E., Sanderson, P. M., & Cameron, I. T. (2010). Using Cognitive Work Analysis Techniques to Identify Human Factor Hazards. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(4), 269–273. <https://doi.org/10.1177/154193121005400401>
- Hausken, K., Welburn, J. W., & Zhuang, J. (2024). A Review of Attacker–Defender Games and Cyber Security. *Games*, 15(4), 28. <https://doi.org/10.3390/g15040028>
- He, F., & Burns, C. M. (2022). A Battle of Voices: A Study of the Relationship between Driving Experience, Driving Style, and In-Vehicle Voice Assistant Character. *Proceedings of the 14th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 236–242. <https://doi.org/10.1145/3543174.3546845>
- Hofbauer, J., Gomez Buquerin, K. K., & Hof, H.-J. (2023). *From SOC to VSOC: Transferring key requirements for efficient vehicle security operations* (p. 15 pages) [Application/pdf]. Ruhr-Universität Bochum. <https://doi.org/10.13154/294-10389>
- Howell, D. C. (2013). *Statistical methods for psychology* (8. ed). Wadsworth, Cengage Learning.
- Hulme, A., McLean, S., Read, G. J. M., Dallat, C., Bedford, A., & Salmon, P. M. (2019). Sports Organizations as Complex Systems: Using Cognitive Work Analysis to Identify the Factors Influencing Performance in an Elite Netball Organization. *Frontiers in Sports and Active Living*, 1, 56. <https://doi.org/10.3389/fspor.2019.00056>
- Hungund, A. P., Pai, G., & Pradhan, A. K. (2021). Systematic Review of Research on Driver Distraction in the Context of Advanced Driver Assistance Systems. *Transportation Research*

- Record: Journal of the Transportation Research Board*, 2675(9), 756–765.  
<https://doi.org/10.1177/03611981211004129>
- Illiashenko, O., Kharchenko, V., Babeshko, I., Fesenko, H., & Di Giandomenico, F. (2023). Security-Informed Safety Analysis of Autonomous Transport Systems Considering AI-Powered Cyberattacks and Protection. *Entropy*, 25(8), 1123. <https://doi.org/10.3390/e25081123>
- Jaeger, L., & Eckhardt, A. (2021). Eyes wide open: The role of situational information security awareness for security-related behaviour. *Information Systems Journal*, 31(3), 429–472. <https://doi.org/10.1111/isj.12317>
- Jalalvand, F., Baruwal Chhetri, M., Nepal, S., & Paris, C. (2025). Alert Prioritisation in Security Operations Centres: A Systematic Survey on Criteria and Methods. *ACM Computing Surveys*, 57(2), 1–36. <https://doi.org/10.1145/3695462>
- Jastrzemski, T. S., & Charness, N. (2007). The Model Human Processor and the older adult: Parameter estimation and validation within a mobile phone task. *Journal of Experimental Psychology: Applied*, 13(4), 224–248. <https://doi.org/10.1037/1076-898X.13.4.224>
- Jenkins, D. P., Stanton, N. A., Salmon, P. M., Walker, G. H., Young, M. S., Whitworth, I., Farmilo, A., & Hone, G. (2007). The Development of a Cognitive Work Analysis Tool. In D. Harris (Ed.), *Engineering Psychology and Cognitive Ergonomics* (Vol. 4562, pp. 504–511). Springer Berlin Heidelberg. [https://doi.org/10.1007/978-3-540-73331-7\\_55](https://doi.org/10.1007/978-3-540-73331-7_55)
- John, B. G. (1998). *Cognitive Modeling for Human-Computer Interaction*.
- Johnson, T. (2017). *A recreation of the Tesla Model 3 UI*.
- Joint Task Force Transformation Initiative. (2012). *Guide for conducting risk assessments* (NIST SP 800-30r1; 0 ed., p. NIST SP 800-30r1). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-30r1>

- Jongen, S., Vuurman, E. F. P. M., Ramaekers, J. G., & Vermeeren, A. (2016). The sensitivity of laboratory tests assessing driving related skills to dose-related impairment of alcohol: A literature review. *Accident Analysis & Prevention*, *89*, 31–48.  
<https://doi.org/10.1016/j.aap.2016.01.001>
- Joshi, H., & Ustun, V. (2024). Augmenting Cognitive Architectures with Large Language Models. *Proceedings of the AAAI Symposium Series*, *2*(1), 281–285.  
<https://doi.org/10.1609/aaaiss.v2i1.27689>
- Jovanovic, M. (2024). *Towards Incremental Learning In Large Language Models A Critical Review*.
- Karantzas, G., & Patsakis, C. (2021). An Empirical Assessment of Endpoint Detection and Response Systems against Advanced Persistent Threats Attack Vectors. *Journal of Cybersecurity and Privacy*, *1*(3), 387–421. <https://doi.org/10.3390/jcp1030021>
- Katrakazas, C., Theofilatos, A., Papastefanatos, G., Härrä, J., & Antoniou, C. (2020). Cyber security and its impact on CAV safety: Overview, policy needs and challenges. In *Advances in Transport Policy and Planning* (Vol. 5, pp. 73–94). Elsevier.  
<https://doi.org/10.1016/bs.atpp.2020.05.001>
- Kelley, T. D. (2003). Symbolic and Sub-Symbolic Representations in Computational Models of Human Cognition: What Can be Learned from Biology? *Theory & Psychology*, *13*(6), 847–860. <https://doi.org/10.1177/0959354303136005>
- Kennedy, R., Clifford, S., Burleigh, T., Waggoner, P. D., Jewell, R., & Winter, N. J. G. (2020). The shape of and solutions to the MTurk quality crisis. *Political Science Research and Methods*, *8*(4), 614–629. <https://doi.org/10.1017/psrm.2020.6>
- Kersten, L., Darré, S., Mulders, T., Zambon, E., Caselli, M., Snijders, C., & Allodi, L. (2024). A Security Alert Investigation Tool Supporting Tier 1 Analysts in Contextualizing and

- Understanding Network Security Events. *2024 Annual Computer Security Applications Conference (ACSAC)*, 890–905. <https://doi.org/10.1109/ACSAC63791.2024.00076>
- Kersten, L., Mulders, T., Zambon, E., Snijders, C., & Allodi, L. (n.d.). 'Give Me Structure': *Synthesis and Evaluation of a (Network) Threat Analysis Process Supporting Tier 1 Investigations in a Security Operation Center*.
- Khadka, K., & Ullah, A. B. (2025). Human factors in cybersecurity: An interdisciplinary review and framework proposal. *International Journal of Information Security*, 24(3), 119. <https://doi.org/10.1007/s10207-025-01032-0>
- Khan, S. K., Shiwakoti, N., Stasinopoulos, P., & Chen, Y. (2020). Cyber-attacks in the next-generation cars, mitigation techniques, anticipated readiness and future directions. *Accident Analysis & Prevention*, 148, 105837. <https://doi.org/10.1016/j.aap.2020.105837>
- Khan, S. K., Shiwakoti, N., Stasinopoulos, P., Chen, Y., & Warren, M. (2025). Cybersecurity framework for connected and automated vehicles: A modelling perspective. *Transport Policy*, 162, 47–64. <https://doi.org/10.1016/j.tranpol.2024.11.019>
- Kidmose, B. (2025). A review of smart vehicles in smart cities: Dangers, impacts, and the threat landscape. *Vehicular Communications*, 51, 100871. <https://doi.org/10.1016/j.vehcom.2024.100871>
- Kieras, D., Meyer, D., & Ballas, J. (2001). Towards demystification of direct manipulation: Cognitive modeling charts the gulf of execution. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 128–135. <https://doi.org/10.1145/365024.365069>
- Kim, K., Kim, J. S., Jeong, S., Park, J.-H., & Kim, H. K. (2021). Cybersecurity for autonomous vehicles: Review of attacks and defense. *Computers & Security*, 103, 102150. <https://doi.org/10.1016/j.cose.2020.102150>

- Kim, S., & Shrestha, R. (2020). *Automotive Cyber Security: Introduction, Challenges, and Standardization*. Springer Singapore. <https://doi.org/10.1007/978-981-15-8053-6>
- Knerler, K., Parker, I., & Zimmerman, C. (2023). 11 Strategies of a World-Class Cybersecurity Operations Center: Highlights. *MITRE*.
- Knight, I. A., Wilson, M., Brailsford, D., & Natasa Milic-Frayling. (2018). “*Enslaved to the Trapped Data*”: A Cognitive Work Analysis of Medical Systematic Reviews. <https://doi.org/10.13140/RG.2.2.10311.75684/1>
- Koehler, D. J., & Harvey, N. (2008). *Blackwell Handbook of Judgment and Decision Making*. John Wiley & Sons.
- Kokulu, F. B., Soneji, A., Bao, T., Shoshitaishvili, Y., Zhao, Z., Doupé, A., & Ahn, G.-J. (2019). Matched and Mismatched SOCs: A Qualitative Study on Security Operations Center Issues. *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, 1955–1970. <https://doi.org/10.1145/3319535.3354239>
- Kotseruba, I., & Tsotsos, J. K. (2016). *A Review of 40 Years of Cognitive Architecture Research: Core Cognitive Abilities and Practical Applications* (Version 3). arXiv. <https://doi.org/10.48550/ARXIV.1610.08602>
- Kotseruba, I., & Tsotsos, J. K. (2020). 40 years of cognitive architectures: Core cognitive abilities and practical applications. *Artificial Intelligence Review*, 53(1), 17–94. <https://doi.org/10.1007/s10462-018-9646-y>
- Kovesdi, C. R., & Spielman, Z. A. (2021). Exploring the Use of Cognitive Work Analysis in Developing a Nuclear Power Plant New-State Vision. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1), 452–456. <https://doi.org/10.1177/1071181321651129>

- Krause, T., Ernst, R., Klaer, B., Hacker, I., & Henze, M. (2021). Cybersecurity in Power Grids: Challenges and Opportunities. *Sensors*, 21(18), 6225. <https://doi.org/10.3390/s21186225>
- Laban, P., Hayashi, H., Zhou, Y., & Neville, J. (2025). *LLMs Get Lost In Multi-Turn Conversation* (arXiv:2505.06120). arXiv. <https://doi.org/10.48550/arXiv.2505.06120>
- Laird, J. E. (2021). *An Analysis and Comparison of ACT-R and Soar*.
- Laird, J. E., & Congdon, C. B. (2004). *The Soar User's Manual Version 8.5 Edition 1*.  
[http://sitemaker.umich.edu/soar/soar\\_software\\_downloads](http://sitemaker.umich.edu/soar/soar_software_downloads)
- Langley, P. (2006). *A Unified Cognitive Architecture for Physical Agents*.
- Langley, P., Laird, J. E., & Rogers, S. (2009). Cognitive architectures: Research issues and challenges. *Cognitive Systems Research*, 10(2), 141–160.  
<https://doi.org/10.1016/j.cogsys.2006.07.004>
- Lebière, C., Anderson, J. R., & Reder, L. M. (2019). Error Modeling in the ACT-R Production System. In A. Ram & K. Eiselt (Eds.), *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (1st ed., pp. 555–559). Routledge.  
<https://doi.org/10.4324/9781315789354-96>
- Lehman, J. F., Laird, J., & Rosenbloom, P. (1996). *A Gentle Introduction to Soar, an Architecture for Human Cognition*.
- Li, X., Zhao, J., Cao, Y., Liu, J., Yan, J., & Li, H. (2023). A Framework Design to Improve the Operation Capability of Cyber Security for Electric Power Dispatching and Control Systems based on the Concept of SOC. *Journal of Physics: Conference Series*, 2418(1), 012074.  
<https://doi.org/10.1088/1742-6596/2418/1/012074>
- Li, Z., Jin, D., Hannon, C., Shahidehpour, M., & Wang, J. (2016). Assessing and mitigating cybersecurity risks of traffic light systems in smart cities. *IET Cyber-Physical Systems: Theory & Applications*, 1(1), 60–69. <https://doi.org/10.1049/iet-cps.2016.0017>

- Li, Z., Zhu, H., Lu, Z., Xiao, Z., & Yin, M. (2025). From Text to Trust: Empowering AI-assisted Decision Making with Adaptive LLM-powered Analysis. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, 1–18.  
<https://doi.org/10.1145/3706598.3713133>
- Linkov, V., Zámečník, P., Havlíčková, D., & Pai, C.-W. (2019). Human Factors in the Cybersecurity of Autonomous Vehicles: Trends in Current Research. *Frontiers in Psychology*, *10*, 995.  
<https://doi.org/10.3389/fpsyg.2019.00995>
- Litherland, P., Orr, R., & Piggin, R. (2016). Cyber security of operational technology: Understanding differences and achieving balance between nuclear safety and nuclear security. *11th International Conference on System Safety and Cyber-Security (SSCS 2016)*, 6 .-6 .  
<https://doi.org/10.1049/cp.2016.0856>
- Liu, H., Ning, R., Teng, Z., Liu, J., Zhou, Q., & Zhang, Y. (2023). *Evaluating the Logical Reasoning Ability of ChatGPT and GPT-4* (arXiv:2304.03439). arXiv.  
<https://doi.org/10.48550/arXiv.2304.03439>
- Liu, N., Sonkar, S., Wang, Z., Woodhead, S., & Baraniuk, R. G. (2023). *Novice Learner and Expert Tutor: Evaluating Math Reasoning Abilities of Large Language Models with Misconceptions* (arXiv:2310.02439). arXiv. <https://doi.org/10.48550/arXiv.2310.02439>
- Liu, R., Geng, J., Wu, A. J., Sucholutsky, I., Lombrozo, T., & Griffiths, T. L. (2025). *Mind Your Step (by Step): Chain-of-Thought can Reduce Performance on Tasks where Thinking Makes Humans Worse* (arXiv:2410.21333). arXiv. <https://doi.org/10.48550/arXiv.2410.21333>
- Liu, Y., Feyen, R., & Tsimhoni, O. (2006). Queueing Network-Model Human Processor (QN-MHP): A computational architecture for multitask performance in human-machine systems. *ACM Transactions on Computer-Human Interaction*, *13*(1), 37–70.  
<https://doi.org/10.1145/1143518.1143520>

- Lombrozo, T. (2020). "Learning by Thinking" in Science and in Everyday Life. In A. Levy & P. Godfrey-Smith (Eds.), *The Scientific Imagination* (1st ed., pp. 230–249). Oxford University Press New York. <https://doi.org/10.1093/oso/9780190212308.003.0010>
- Lopez, A., Malawade, A. V., Al Faruque, M. A., Boddupalli, S., & Ray, S. (2019). Security of Emergent Automotive Systems: A Tutorial Introduction and Perspectives on Practice. *IEEE Design & Test*, 36(6), 10–38. <https://doi.org/10.1109/MDAT.2019.2944086>
- Ludwig, J. (2005). *Psychologically Inspired Symbolic Cognitive Architectures*.
- Luo, Y., Yu, X., Tran, T. T. M., & Hoggenmueller, M. (2025). Uncertainty on Display: The Effects of Communicating Confidence Cues in Autonomous Vehicle-Pedestrian Interactions. *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, 307–317. <https://doi.org/10.1145/3744333.3747826>
- Maalem Lahcen, R. A., Caulkins, B., Mohapatra, R., & Kumar, M. (2020). Review and insight on the behavioral aspects of cybersecurity. *Cybersecurity*, 3(1), 10. <https://doi.org/10.1186/s42400-020-00050-w>
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., Alon, U., Dziri, N., Prabhumoye, S., Yang, Y., Gupta, S., Majumder, B. P., Hermann, K., Welleck, S., Yazdanbakhsh, A., & Clark, P. (2023). *SELF-REFINE: Iterative Refinement with Self-Feedback*.
- Malik, S., & Sun, W. (2020). Analysis and Simulation of Cyber Attacks Against Connected and Autonomous Vehicles. *2020 International Conference on Connected and Autonomous Driving (MetroCAD)*, 62–70. <https://doi.org/10.1109/MetroCAD48866.2020.00018>
- Malloy, T., & Gonzalez, C. (2024). Applying Generative Artificial Intelligence to cognitive models of decision making. *Frontiers in Psychology*, 15, 1387948. <https://doi.org/10.3389/fpsyg.2024.1387948>

- Marewski, J. N., & Mehlhorn, K. (2011). Using the ACT-R architecture to specify 39 quantitative process models of decision making. *Judgment and Decision Making*, 6(6), 439–519.  
<https://doi.org/10.1017/s1930297500002473>
- Martínez-Cruz, A., Ramírez-Gutiérrez, K. A., Feregrino-Uribe, C., & Morales-Reyes, A. (2021). Security on in-vehicle communication protocols: Issues, challenges, and future research directions. *Computer Communications*, 180, 1–20.  
<https://doi.org/10.1016/j.comcom.2021.08.027>
- Martín-Pérez, M., Marias Parella, J., Fernández, J., De Juan Fidalgo, P., Casademont, J., Álvarez Romero, A., & Diaz-Rodriguez, R. (2023). A Testbed for a Nearby-Context Aware: Threat Detection and Mitigation System for Connected Vehicles. *2023 JNIC Cybersecurity Conference (JNIC)*, 1–8. <https://doi.org/10.23919/JNIC58574.2023.10205891>
- McClelland, J. L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, 1(1), 11–38. <https://doi.org/10.1111/j.1756-8765.2008.01003.x>
- Meyers, C., Powers, S., & Faissol, D. (2009). *Taxonomies of Cyber Adversaries and Attacks: A Survey of Incidents and Approaches* (LLNL-TR-419041, 967712; p. LLNL-TR-419041, 967712). <https://doi.org/10.2172/967712>
- Michalec, O., Shreeve, B., & Rashid, A. (2023). Who will keep the lights on? Expertise and inclusion in cyber security visions of future energy systems. *Energy Research & Social Science*, 106, 103327. <https://doi.org/10.1016/j.erss.2023.103327>
- Mitchell, D. K. (2003). *Advanced Improved Performance Research Integration Tool (IMPRINT) Vetrionics Technology Test Bed Model Development*. US Dept of the Army.  
<https://doi.org/10.21236/ada417350>
- Mozannar, H., Bansal, G., Tan, C., Fourney, A., Dibia, V., Niedtner, F., Gerrits, J., Alber, J., Chen, J., Bassman, G., Zhu, E. (Eric), Chang, P., Loynd, R., Murad, M., Hosn, R., Kamar, E., &

- Amershi, S. (2025, May). *Magentic-UI, an experimental human-centered web agent*.  
<https://www.microsoft.com/en-us/research/blog/magentic-ui-an-experimental-human-centered-web-agent/>
- Muniz, J., AlFardan, N., & McIntyre, G. (2015). *Security operations center: Building, operating, and maintaining your SOC* (1st ed.). Cisco Press.
- Mutzenich, C., Durant, S., Helman, S., & Dalton, P. (2021). Updating our understanding of situation awareness in relation to remote operators of autonomous vehicles. *Cognitive Research: Principles and Implications*, 6(1), 9. <https://doi.org/10.1186/s41235-021-00271-8>
- Muzahid, A. J. M., Zhao, X., & Wang, Z. (2024). *Survey on Human-Vehicle Interactions and AI Collaboration for Optimal Decision-Making in Automated Driving* (Version 1). arXiv.  
<https://doi.org/10.48550/ARXIV.2412.08005>
- Mwanje, M. D., Kaiwartya, O., Aljaidi, M., Cao, Y., Kumar, S., Jha, D. N., Naser, A., & Lloret, J. (2024). Cyber security analysis of connected vehicles. *IET Intelligent Transport Systems*, 18(7), 1175–1195. <https://doi.org/10.1049/itr2.12504>
- Naikar, N. (2017). Cognitive work analysis: An influential legacy extending beyond human factors and engineering. *Applied Ergonomics*, 59, 528–540.  
<https://doi.org/10.1016/j.apergo.2016.06.001>
- Newell, A. (1994). *Unified theories of cognition* (3. print., 1. Harvard Univ. Press paperback ed). Harvard Univ. Press.
- Newhouse, W., Keith, S., Scribner, B., & Witte, G. (2017). *National Initiative for Cybersecurity Education (NICE) Cybersecurity Workforce Framework* (NIST SP 800-181; p. NIST SP 800-181). National Institute of Standards and Technology. <https://doi.org/10.6028/NIST.SP.800-181>

- Nguyen, T. N., Jamale, K., & Gonzalez, C. (2024). Predicting and Understanding Human Action Decisions: Insights from Large Language Models and Cognitive Instance-Based Learning. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 12*, 126–136. <https://doi.org/10.1609/hcomp.v12i1.31607>
- NHTSA. (2021). *Vehicle Cybersecurity*. <https://www.nhtsa.gov/research/vehicle-cybersecurity>
- Nikitas, A., Parkinson, S., & Vallati, M. (2022). The deceitful Connected and Autonomous Vehicle: Defining the concept, contextualising its dimensions and proposing mitigation policies. *Transport Policy, 122*, 1–10. <https://doi.org/10.1016/j.tranpol.2022.04.011>
- Niu, Q., Liu, J., Bi, Z., Feng, P., Peng, B., Chen, K., Li, M., Yan, L. K., Zhang, Y., Yin, C. H., Fei, C., Wang, T., Wang, Y., Chen, S., & Liu, M. (2024). *Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges* (arXiv:2409.02387). arXiv. <https://doi.org/10.48550/arXiv.2409.02387>
- Nobles, C. (2022). Stress, Burnout, and Security Fatigue in Cybersecurity: A Human Factors Problem. *HOLISTICA – Journal of Business and Public Administration, 13*(1), 49–72. <https://doi.org/10.2478/hjbpa-2022-0003>
- Nowdehi, N., Aoudi, W., Almgren, M., & Olovsson, T. (2019). *CASAD: CAN-Aware Stealthy-Attack Detection for In-Vehicle Networks* (arXiv:1909.08407). arXiv. <https://doi.org/10.48550/arXiv.1909.08407>
- Noy, I. Y., Shinar, D., & Horrey, W. J. (2018). Automated driving: Safety blind spots. *Safety Science, 102*, 68–78. <https://doi.org/10.1016/j.ssci.2017.07.018>
- Nyamsuren, E., & Taatgen, N. A. (2013). Pre-attentive and attentive vision module. *Cognitive Systems Research, 24*, 62–71. <https://doi.org/10.1016/j.cogsys.2012.12.010>
- Ofte, H. J., & Katsikas, S. (2023). Understanding situation awareness in SOCs, a systematic literature review. *Computers & Security, 126*, 103069. <https://doi.org/10.1016/j.cose.2022.103069>

- Oladimeji, D., Gupta, K., Kose, N. A., Gundogan, K., Ge, L., & Liang, F. (2023). Smart Transportation: An Overview of Technologies and Applications. *Sensors*, 23(8), 3880. <https://doi.org/10.3390/s23083880>
- Olt, C. (2019). Establishing Security Operation Centers for Connected Cars. *ATZelectronics Worldwide*, 14(5), 40–43. <https://doi.org/10.1007/s38314-019-0050-4>
- Oniagbi, O. (2019). *Evaluation of LLM Agents for the SOC Tier 1 Analyst Triage Process*.
- Pan, X., Lin, Y., & He, C. (2017). A Review of Cognitive Models in Human Reliability Analysis. *Quality and Reliability Engineering International*, 33(7), 1299–1316. <https://doi.org/10.1002/qre.2111>
- Paolacci, G., & Chandler, J. (2014). Inside the Turk: Understanding Mechanical Turk as a Participant Pool. *Current Directions in Psychological Science*, 23(3), 184–188. <https://doi.org/10.1177/0963721414531598>
- Park, J., & Zahabi, M. (2024). A Review of Human Performance Models for Prediction of Driver Behavior and Interactions With In-Vehicle Technology. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 66(4), 1249–1275. <https://doi.org/10.1177/00187208221132740>
- Parker, S., Wu, Z., & Christofides, P. D. (2023). Cybersecurity in process control, operations, and supply chain. *Computers & Chemical Engineering*, 171, 108169. <https://doi.org/10.1016/j.compchemeng.2023.108169>
- Passi, S., & Vorvoreanu, M. (2022). *Overreliance on AI Literature Review*.
- Payre, W., Perelló-March, J., Kanakapura Sriranga, A., & Birrell, S. (2023). The notorious B.I.T: The effects of a ransomware and a screen failure on distraction in automated driving. *Transportation Research Part F: Traffic Psychology and Behaviour*, 94, 42–52. <https://doi.org/10.1016/j.trf.2023.02.002>

- Petersen, R., Santos, D., Smith, M. C., Wetzel, K. A., & Witte, G. (2020). *Workforce Framework for Cybersecurity (NICE Framework)*. National Institute of Standards and Technology.  
<https://doi.org/10.6028/NIST.SP.800-181r1>
- Petit, J., & Shladover, S. E. (2014). Potential Cyberattacks on Automated Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 1–11.  
<https://doi.org/10.1109/TITS.2014.2342271>
- Pew, R. W., Mavor, A. S., & National Research Council (U.S.) (Eds.). (1998). *Modeling human and organizational behavior: Application to military simulations*. National Academy Press.
- Pham, M., & Xiong, K. (2020). *A Survey on Security Attacks and Defense Techniques for Connected and Autonomous Vehicles* (arXiv:2007.08041). arXiv.  
<https://doi.org/10.48550/arXiv.2007.08041>
- Prebot, B., Du, Y., & Xi, X. (2022). *Cognitive Models of Dynamic Decisions in Autonomous Intelligent Cyber Defense*.
- Ralethe, S., & Buys, J. (2022). *Generic Overgeneralization in Pre-trained Language Models*.
- Ranade, P., Piplai, A., Joshi, A., & Finin, T. (2021). CyBERT: Contextualized Embeddings for the Cybersecurity Domain. *2021 IEEE International Conference on Big Data (Big Data)*, 3334–3342. <https://doi.org/10.1109/BigData52589.2021.9671824>
- Rasmussen, J. (1983). Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-13*(3), 257–266. <https://doi.org/10.1109/TSMC.1983.6313160>
- Rasmussen, J. (1985). The role of hierarchical knowledge representation in decisionmaking and system management. *IEEE Transactions on Systems, Man, and Cybernetics, SMC-15*(2), 234–243. <https://doi.org/10.1109/TSMC.1985.6313353>

- Rasmussen, J., & Jensen, A. (1974). Mental Procedures in Real-Life Tasks: A Case Study of Electronic Trouble Shooting. *Ergonomics*, *17*(3), 293–307.  
<https://doi.org/10.1080/00140137408931355>
- Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering*. Wiley.
- Rauffet, P., Chauvin, C., Morel, G., & Berruet, P. (2015). Designing sociotechnical systems: A CWA-based method for dynamic function allocation. *Proceedings of the European Conference on Cognitive Ergonomics 2015*, 1–8. <https://doi.org/10.1145/2788412.2788433>
- Reed, G. F., Lynn, F., & Meade, B. D. (2002). Use of Coefficient of Variation in Assessing Variability of Quantitative Assays. *Clinical and Vaccine Immunology*, *9*(6), 1235–1239.  
<https://doi.org/10.1128/CDLI.9.6.1235-1239.2002>
- Rehman, U., Cao, S., & Macgregor, C. G. (2024). Modeling Brake Perception Response Time in On-Road and Roadside Hazards Using an Integrated Cognitive Architecture. *IEEE Transactions on Human-Machine Systems*, *54*(4), 441–454. <https://doi.org/10.1109/THMS.2024.3408841>
- Ritter, F. E. (2019). Modeling human cognitive behavior for system design. In *DHM and Posturography* (pp. 517–525). Elsevier. <https://doi.org/10.1016/B978-0-12-816713-7.00037-4>
- Ritter, F. E., Shadbolt, N. R., Elliman, D., Young, R. M., Gobet, F., & Baxter, G. D. (2003). *Techniques for Modeling Human Performance in Synthetic Environments: A Supplementary Review*. OSD or Non-Service DoD Agency. <https://doi.org/10.21236/ada487721>
- Rosso, M., Campobasso, M., Gankhuyag, G., & Allodi, L. (2022). SAIBERSOC: A Methodology and Tool for Experimenting with Security Operation Centers. *Digital Threats: Research and Practice*, *3*(2), 1–29. <https://doi.org/10.1145/3491266>
- Rudd, E. M., Rozsa, A., Gunther, M., & Boulton, T. E. (2017). A Survey of Stealth Malware Attacks, Mitigation Measures, and Steps Toward Autonomous Open World Solutions. *IEEE*

- Communications Surveys & Tutorials*, 19(2), 1145–1172.  
<https://doi.org/10.1109/COMST.2016.2636078>
- Russell, S., & Norvig, P. (2022). *Artificial Intelligence A Modern Approach*.
- Sadaf, M., Iqbal, Z., Javed, A. R., Saba, I., Krichen, M., Majeed, S., & Raza, A. (2023). Connected and Automated Vehicles: Infrastructure, Applications, Security, Critical Challenges, and Future Aspects. *Technologies*, 11(5), 117. <https://doi.org/10.3390/technologies11050117>
- SAE International. (2018). *SURFACE VEHICLE RECOMMENDED PRACTICE (J3016): Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.
- SAE International. (2021a). *SAE J3016\_202104: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*. SAE International.  
[https://www.sae.org/standards/content/j3016\\_202104/](https://www.sae.org/standards/content/j3016_202104/)
- SAE International. (2021b). *SURFACE VEHICLE RECOMMENDED PRACTICE (J3016): Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*.
- Salfati, E., & Pease, M. (2022). *Digital Forensics and Incident Response (DFIR) framework for Operational Technology (OT)* (NIST IR 8428; p. NIST IR 8428). National Institute of Standards and Technology (U.S.). <https://doi.org/10.6028/NIST.IR.8428>
- Salmon, P. M., Lenné, M. G., Read, G. J. M., Mulvihill, C. M., Cornelissen, M., Young, K. L., Walker, G. H., Stanton, N. A., & Stevens, N. (2015). Beyond the Crossing: A Cognitive Work Analysis of Rail Level Crossing Systems. *Procedia Manufacturing*, 3, 2921–2928.  
<https://doi.org/10.1016/j.promfg.2015.07.818>
- Salmon, P. M., Read, G. J. M., Stevens, N., Walker, G. H., Beanland, V., McClure, R., Hughes, B., Johnston, I. R., & Stanton, N. A. (2019). Using the abstraction hierarchy to identify how the

- purpose and structure of road transport systems contributes to road trauma. *Transportation Research Interdisciplinary Perspectives*, 3, 100067.  
<https://doi.org/10.1016/j.trip.2019.100067>
- Salmon, P. M., Regan, M., Lenne, M. G., Stanton, N. A., & Young, K. (2007). Work domain analysis and intelligent transport systems: Implications for vehicle design. *International Journal of Vehicle Design*, 45(3), 426. <https://doi.org/10.1504/IJVD.2007.014914>
- Salvucci, D. D. (2009). Rapid prototyping and evaluation of in-vehicle interfaces. *ACM Transactions on Computer-Human Interaction*, 16(2), 1–33. <https://doi.org/10.1145/1534903.1534906>
- Salvucci, D. D., Chavez, A. K., & Lee, F. J. (2004). *Modeling Effects of Age in Complex Tasks: A Case Study in Driving*.
- Salvucci, D. D., & Taatgen, N. A. (2008). Threaded cognition: An integrated theory of concurrent multitasking. *Psychological Review*, 115(1), 101–130. <https://doi.org/10.1037/0033-295X.115.1.101>
- Sample, C., Hutchinson, S., Maple, C., & Karmanian, A. (2018, April). *Cultural Observations on Social Engineering Victims*. European Conference on Cyber Warfare and Security, Dublin, Ireland.
- Saulaiman, M. N.-E., Bánáti, A., Kozlovsky, M., Kövesi, K. Z., Pozsonyi, T. G., & Csilling, Á. (2025). Cloud-Based Cybersecurity and Data Management System for Near Real-Time Monitoring and Alerting in Vehicle-SOC - a Proof of Concept. *2025 IEEE 23rd World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 000165–000170. <https://doi.org/10.1109/SAMI63904.2025.10883171>
- Sawyer, B. D., & Hancock, P. A. (2018). Hacking the Human: The Prevalence Paradox in Cybersecurity. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 60(5), 597–609. <https://doi.org/10.1177/0018720818780472>

- Schmid, D., Korn, B., & Stanton, N. A. (2020). Evaluating the reduced flight deck crew concept using cognitive work analysis and social network analysis: Comparing normal and data-link outage scenarios. *Cognition, Technology & Work*, 22(1), 109–124.  
<https://doi.org/10.1007/s10111-019-00548-5>
- Seppelt, B. D., & Lee, J. D. (2007). Making adaptive cruise control (ACC) limits visible. *Computer Studies*.
- Shahjee, D., & Ware, N. (2022). Designing a Framework of an Integrated Network and Security Operation Center: A Convergence Approach. *2022 IEEE 7th International Conference for Convergence in Technology (I2CT)*, 1–4. <https://doi.org/10.1109/I2CT54291.2022.9825084>
- Sheehan, B., Murphy, F., Mullins, M., & Ryan, C. (2019). Connected and autonomous vehicles: A cyber-risk classification framework. *Transportation Research Part A: Policy and Practice*, 124, 523–536. <https://doi.org/10.1016/j.tra.2018.06.033>
- Song, H. (2021). The Necessity Analysis of Car Center Console in Car Based on Virtual Technology. *2021 International Conference on Computer, Internet of Things and Control Engineering (CITCE)*, 171–176. <https://doi.org/10.1109/CITCE54390.2021.00041>
- Stanton, N. A., & Allison, C. K. (2020). Driving towards a greener future: An application of cognitive work analysis to promote fuel-efficient driving. *Cognition, Technology & Work*, 22(1), 125–142. <https://doi.org/10.1007/s10111-019-00554-7>
- Stewart, H., & Jürjens, J. (2017). Information security management and the human aspect in organizations. *Information & Computer Security*, 25(5), 494–534.  
<https://doi.org/10.1108/ICS-07-2016-0054>
- Still, A. (2016). *A History of Creativity for Future AI Research*.
- Strayer, D. L., Drews, F. A., & Crouch, D. J. (2006). A Comparison of the Cell Phone Driver and the Drunk Driver. *Human Factors*.

- Sun, P., Liu, Y., Wu, G., & Duan, Z. (2020). *Research on fault diagnosis of reactor coolant accident in nuclear power plant based on radial basis function and fuzzy neural network*. 6.
- Sun, R. (2008). *Introduction to Computational Cognitive Modeling*.
- Sun, R., Merrill, E., & Peterson, T. (2001). From implicit skills to explicit knowledge: A bottom-up model of skill learning. *Cognitive Science*, 25(2), 203–244.  
[https://doi.org/10.1207/s15516709cog2502\\_2](https://doi.org/10.1207/s15516709cog2502_2)
- Sun, X., Yu, F. R., & Zhang, P. (2022). A Survey on Cyber-Security of Connected and Autonomous Vehicles (CAVs). *IEEE Transactions on Intelligent Transportation Systems*, 23(7), 6240–6259. <https://doi.org/10.1109/TITS.2021.3085297>
- Taatgen, N. A., & Anderson, J. R. (2002). Why do children learn to say “Broke”? A model of learning the past tense without feedback. *Cognition*, 86(2), 123–155.  
[https://doi.org/10.1016/S0010-0277\(02\)00176-2](https://doi.org/10.1016/S0010-0277(02)00176-2)
- Taatgen, N. A., & Lee, F. J. (2003). Production Compilation: A Simple Mechanism to Model Complex Skill Acquisition. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(1), 61–76. <https://doi.org/10.1518/hfes.45.1.61.27224>
- Taatgen, N., & Anderson, J. (2008). *Running Head: CONSTRAINTS IN COGNITIVE ARCHITECTURES*.
- Taatgen, N., & Anderson, J. R. (2010). The Past, Present, and Future of Cognitive Architectures: Topics in Cognitive Science. *Topics in Cognitive Science*, 2(4), 693–704.  
<https://doi.org/10.1111/j.1756-8765.2009.01063.x>
- Takahashi, T., Kadobayashi, Y., & Naka, K. (2011). *Toward global cybersecurity collaboration: Cybersecurity operation activity model*. 1–8.

- Tariq, S., Baruwal Chhetri, M., Nepal, S., & Paris, C. (2025). Alert Fatigue in Security Operations Centres: Research Challenges and Opportunities. *ACM Computing Surveys*, 57(9), 1–38. <https://doi.org/10.1145/3723158>
- Thackray, H., McAlaney, J., Dogan, H., Taylor, J., & Richardson, C. (2016, July). *Social Psychology: An under-used tool in Cybersecurity*. Proceedings of the 30th International BCS Human Computer Interaction Conference. <https://doi.org/10.14236/ewic/HCI2016.64>
- The ACT-R Cognitive Architecture and Its pyactr Implementation* (with Brasoveanu, A., & Dotlačil, J.). (2020). Springer International Publishing. [https://doi.org/10.1007/978-3-030-31846-8\\_2](https://doi.org/10.1007/978-3-030-31846-8_2)
- The Splunk SOAR Service*. (2023). [Computer software]. Splunk® SOAR.
- Trend Micro Research. (2020). ISO/SAE 21434: Setting the Standard for Connected Cars' Cybersecurity. *White Paper*.
- Tsiknas, K., Taketzis, D., Demertzis, K., & Skianis, C. (2021). Cyber Threats to Industrial IoT: A Survey on Attacks and Countermeasures. *IoT*, 2(1), 163–186. <https://doi.org/10.3390/iot2010009>
- Tulis, M., Steuer, G., & Dresel, M. (2016). Learning from errors: A model of individual processes. *Frontline Learning Research*, 4(4), 12–26. <https://doi.org/10.14786/flr.v4i2.168>
- Tzoannos, Z.-R., Kosmanos, D., Xenakis, A., & Chaikalis, C. (2024). The Impact of Spoofing Attacks in Connected Autonomous Vehicles under Traffic Congestion Conditions. *Telecom*, 5(3), 747–759. <https://doi.org/10.3390/telecom5030037>
- Van Der Kleij, R., Schraagen, J. M., Cadet, B., & Young, H. (2022). Developing decision support for cybersecurity threat and incident managers. *Computers & Security*, 113, 102535. <https://doi.org/10.1016/j.cose.2021.102535>
- Veksler, V. D., Buchler, N., Hoffman, B. E., Cassenti, D. N., Sample, C., & Sugrim, S. (2018). Simulations in Cyber-Security: A Review of Cognitive Modeling of Network Attackers,

- Defenders, and Users. *Frontiers in Psychology*, 9, 691.  
<https://doi.org/10.3389/fpsyg.2018.00691>
- Veres, C. (2022). *A Precip of Language Models are not Models of Language* (Version 1). arXiv.  
<https://doi.org/10.48550/ARXIV.2205.07634>
- Vicente, K. J. (1999). *Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. CRC Press.
- Vielberth, M., Bohm, F., Fichtinger, I., & Pernul, G. (2020). Security Operations Center: A Systematic Study and Open Challenges. *IEEE Access*, 8, 227756–227779.  
<https://doi.org/10.1109/ACCESS.2020.3045514>
- Vivek, S., Yanni, D., Yunker, P. J., & Silverberg, J. L. (2019). Cyberphysical risks of hacked internet-connected vehicles. *Physical Review E*, 100(1), 012316.  
<https://doi.org/10.1103/PhysRevE.100.012316>
- Wang, J., Yan, T., An, D., Liang, Z., Guo, C., Hu, H., Luo, Q., Li, H., Wang, H., Zeng, S., Zhou, C., Ma, L., & Qi, F. (2021). A comprehensive security operation center based on big data analytics and threat intelligence. *Proceedings of International Symposium on Grids & Clouds 2021 — PoS(ISGC2021)*, 028. <https://doi.org/10.22323/1.378.0028>
- Wang, M., Parker, J., Zhang, F., & Roberts, S. C. (2024). A simulator study assessing the effectiveness of training and warning systems on drivers' response performance to vehicle cyberattacks. *Accident Analysis & Prevention*, 203, 107644.  
<https://doi.org/10.1016/j.aap.2024.107644>
- Wang, R. E., Zhang, Q., Robinson, C., Loeb, S., & Demszky, D. (2024). *Bridging the Novice-Expert Gap via Models of Decision-Making: A Case Study on Remediating Math Mistakes* (arXiv:2310.10648). arXiv. <https://doi.org/10.48550/arXiv.2310.10648>

- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., & Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*.
- Welz, M., & Alfons, A. (2025). *When Respondents Don't Care Anymore: Identifying the Onset of Careless Responding* (arXiv:2303.07167). arXiv. <https://doi.org/10.48550/arXiv.2303.07167>
- Werlinger, R., Muldner, K., Hawkey, K., & Beznosov, K. (2010). Preparation, detection, and analysis: The diagnostic work of IT security incident response. *Information Management & Computer Security*, 18(1), 26–42. <https://doi.org/10.1108/09685221011035241>
- Wu, B., Yip, T. L., Yan, X., & Guedes Soares, C. (2022). Review of techniques and challenges of human and organizational factors analysis in maritime transportation. *Reliability Engineering & System Safety*, 219, 108249. <https://doi.org/10.1016/j.ress.2021.108249>
- Wu, C., Tsimhoni, O., & Liu, Y. (2008). Development of an Adaptive Workload Management System Using the Queueing Network-Model Human Processor (QN-MHP). *IEEE Transactions on Intelligent Transportation Systems*, 9(3), 463–475. <https://doi.org/10.1109/TITS.2008.928172>
- Wurst, C., Chen, H.-Y. W., & Joseph, K. (2021). Formative Modeling of Foster Care Work: A Cognitive Work Analysis Approach. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 65(1), 933–937. <https://doi.org/10.1177/1071181321651023>
- Xie, Y., Zhou, Y., Xu, J., Zhou, J., Chen, X., & Xiao, F. (2021). Cybersecurity protection on in-vehicle networks for distributed automotive cyber-physical systems: State-of-the-art and future challenges. *Software: Practice and Experience*, 51(11), 2108–2127. <https://doi.org/10.1002/spe.2965>
- Xu, L., Guo, L., Wang, X., Ge, P., & Guan, L. (2025). Predicting Drivers' situation awareness and response times in the emergency situation using an integrated cognitive architecture.

- Transportation Research Part F: Traffic Psychology and Behaviour*, 114, 873–887.  
<https://doi.org/10.1016/j.trf.2025.07.007>
- Xu, R., Cao, S., Kearns, S. K., Niechwiej-Szwedo, E., & Irving, E. (2024). Computational Cognitive Modeling of Pilot Performance in Pre-flight and Take-off Procedures. *Journal of Aviation/Aerospace Education & Research*, 33(4). <https://doi.org/10.58940/2329-258X.2026>
- Yadav, A. K., & Velaga, N. R. (2019). Modelling the relationship between different Blood Alcohol Concentrations and reaction time of young and mature drivers. *Transportation Research Part F: Traffic Psychology and Behaviour*, 64, 227–245. <https://doi.org/10.1016/j.trf.2019.05.011>
- Yağdereli, E., Gemci, C., & Aktaş, A. Z. (2015). A study on cyber-security of autonomous and unmanned vehicles. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology*, 12(4), 369–381. <https://doi.org/10.1177/1548512915575803>
- Yan, Z., Robertson, T., Yan, R., Park, S. Y., Bordoff, S., Chen, Q., & Sprissler, E. (2018). Finding the weakest links in the weakest link: How well do undergraduate students make cybersecurity judgment? *Computers in Human Behavior*, 84, 375–382.  
<https://doi.org/10.1016/j.chb.2018.02.019>
- Yang, F., Zhao, P., Wang, Z., Wang, L., Qiao, B., Zhang, J., Garg, M., Lin, Q., Rajmohan, S., & Zhang, D. (2023). Empower Large Language Model to Perform Better on Industrial Domain-Specific Question Answering. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 294–312.  
<https://doi.org/10.18653/v1/2023.emnlp-industry.29>
- Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). *Tree of Thoughts: Deliberate Problem Solving with Large Language Models* (arXiv:2305.10601). arXiv. <https://doi.org/10.48550/arXiv.2305.10601>

- Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). *ReAct: Synergizing Reasoning and Acting in Language Models* (arXiv:2210.03629). arXiv.  
<https://doi.org/10.48550/arXiv.2210.03629>
- Zhai, S., Wang, L., & Liu, P. (2024). Not in Control, but Liable? Attributing Human Responsibility for Fully Automated Vehicle Accidents. *Engineering*, 33, 121–132.  
<https://doi.org/10.1016/j.eng.2023.10.008>
- Zhang, Y., & Lintern, G. (2024). Work domain modeling of human-automation interaction for in-vehicle automation. *Cognition, Technology & Work*, 26(4), 585–601.  
<https://doi.org/10.1007/s10111-024-00780-8>
- Zhao, J., Wu, M., Zhou, L., Wang, X., & Jia, J. (2022). Cognitive psychology-based artificial intelligence review. *Frontiers in Neuroscience*, 16, 1024316.  
<https://doi.org/10.3389/fnins.2022.1024316>
- Zhong, C., Yen, J., Liu, P., & Erbacher, R. F. (2016). Automate Cybersecurity Data Triage by Leveraging Human Analysts' Cognitive Process. *2016 IEEE 2nd International Conference on Big Data Security on Cloud (BigDataSecurity), IEEE International Conference on High Performance and Smart Computing (HPSC), and IEEE International Conference on Intelligent Data and Security (IDS)*, 357–363. <https://doi.org/10.1109/BigDataSecurity-HPSC-IDS.2016.41>
- Zhong, Q., Zhi, J., Xu, Y., Gao, P., & Feng, S. (2024). Assessing driver distraction from in-vehicle information system: An on-road study exploring the effects of input modalities and secondary task types. *Scientific Reports*, 14(1), 20289. <https://doi.org/10.1038/s41598-024-71226-4>
- Zhou, X., Schmedding, A., Ren, H., Yang, L., Schowitz, P., Smirni, E., & Alemzadeh, H. (2022). *Strategic Safety-Critical Attacks Against an Advanced Driver Assistance System* (arXiv:2204.06768). arXiv. <https://doi.org/10.48550/arXiv.2204.06768>



## Appendix A

### **An Example of Simulating Distraction-Based Cyberattacks Targeting In-Vehicle Human Operations using an ACT-R Based Model**

#### **Informal Interviews for Collecting Participants Insights**

*Participants.* The modeling process began by collecting insights from five participants through a single round of brief, informal discussions. The participants consisted of two with research experiences in ADS, two expert drivers (active drivers with over ten years of driving experience), and a novice driver with recent experience with ‘Full Self-Driving’ (FSD) features (Level 2). They contribute distinct views: expert drivers provide advanced driving skills and expertise, whereas the ADS researchers and novice driver more familiar with ADAS.

The informal interview probed participants’ decision-making processes when facing three specific attack models by asking a general question: "How would you respond if you encountered the anomalies as [Scenario Description]? (See Section for the designed attack scenario details.)"

#### **Evaluation Criteria and Reference Research Settings**

The referenced studies adopted a traffic condition setup involving simple road geometry and low-traffic highway conditions. In Deng et al. (2019a)’s, parked vehicles are placed 233 meters ahead, within visible range when traveling at a speed of 120 km/h. In Yadav and Velaga (2019)’s, the speed limit is set at 177 km/h (110 mph), with a parked vehicle positioned 130 meters ahead as a trigger event in the simulator. While the specific experimental parameters differ slightly between the two studies, the measured effect on reaction times and general driving conditions are comparable. Therefore, we compared the modeling results to a baseline reaction time that excludes any visual or auditory distractions (Gold et al., 2013), which also serves as the baseline reference (3.65s) for the foundational model (Deng, Cao, et al., 2019a), and calculated the alcohol-impaired reaction time reported by Yadav and Velaga (2019), as shown in Table 13.

The Reaction Time defined in Yadav and Velaga (2019)'s work is the perception-reaction time (PRT), which refers to the time taken by the driver to perceive the parked vehicle and place their foot on the brake pedal. Likewise, the Reaction Time defined in our foundational model is the time between perceiving the takeover prompt and the first steering or brake/throttle action observed.

**Table 13.** Reference Values for Baseline Reaction Time (Gold et al., 2013; Deng, Cao, et al., 2019a) and Impaired Performance from BAC Study (Yadav & Velaga, 2019).

Conditions	Baseline	BAC 0.03%	BAC 0.05%	BAC 0.08%
Reaction Time		+64%	+78%	+116%
Reference Values	3.65s	5.99s	6.50s	7.89s

Since our model construction was highly exploratory and the production rules were developed based on informal discussions rather than empirical data, we did not conduct a standard statistical model validation. Instead, we examined whether any simulated results exceeded the effects of BAC impairment reported in prior empirical studies (Yadav & Velaga, 2019), which may indicate a potential severity of driving safety.

## Appendix B

### Study 2 Information Requirements List Based on the SOC AH Model (Full List).

**Table 14.** Information Requirements Based on the SOC AH Model (Alahmadi et al., 2022; Knerler et al., 2023; Newhouse et al., 2017; Oniagbi, 2019).

Level	Variables	Information Requirements	
Functional Purpose	Cyber Defense	Overall Downtime, Overall Threat Response/Recovery time, Threat Mitigation Success Rate, System Vulnerability Index	
	Assets Protection	Availability of Assets, Assets List/Topology	
Generalized Function	Activity Monitoring	Data traffic load and capacity (High, Medium, Low), Resource utilization, Traffic routing information, Network traffic analysis, Endpoint monitoring	
	Suspicious Events Detection	Tool aided red-flag unusual behavior/indicator, unnoticed event correlation	
	Events Communication	Alert/Incident tracking/reporting, collaboration/communication tools availability	
	Incident Characterization	Real-time alerts arrays, detected anomaly prioritization, known IOCs and pattern matching	
	Adversary Investigation & Vulnerability Management	Logs from various sources across the network, Event history, Attack chain visualization	
Physical Function	Alert Management	false positive rate, frequencies of alert types, historical alert tracking (whether similar alerts have occurred before and what actions were taken)	
	Records Keeping & Tracking	Historical records of incident response, suspicious pattern/alert tracking and observations, prevention measures updates, configuration changes, updated IOCs, network and asset status records	
	Data Aggregation & Correlation	Automated correlation of events from various sources to identify attack patterns, impacted scope, Historical data and status, updated configurations	
	Information Validation	timestamps of incidents and rules, shared information on incident response, Historical data and status, updated configurations, communication/collaboration tools	
	Impact & Risk Analysis	Severity level, attack type, alert category, impact scope, recovery cost	
Physical Form	Operational Data	Group 3: User and Device	Asset details (e.g., type of device, operating system, criticality to the business), timestamps, configuration data
		Group 1: Network Details	Logs from firewalls, routers, OSs, servers, and endpoints

			Involved zones, systems, assets information (e.g., what systems exist, their roles, and vulnerabilities).
			Incident status, historical related incidents responses, logs and reports
	Cyber Intelligent Data	Group 2: Threat Indicators and Detections	Attack signature, exploitation methods, IDS/IPS
			Common Vulnerabilities and Exposure (CVE), known IOCs, malicious IP addresses, or domains
			Flagged deviations in user, network, or system activity, anomaly threshold (e.g., deviations & baselines)
	Contextual Data	Group 4: Geolocation and Proxy Information	Geolocation data (location of IP addresses), Geographical data that indicates the origin of connections or traffic
			Constituency organizational structures (e.g., internal/outsourced team, user access level), user behavior patterns (e.g., typical login locations and times for users), authentication events (e.g., login attempts)
Group 5: Payload Analysis		Road traffic conditions, weather conditions, asset details (e.g., type of device, operating system, criticality to the business), etc.	

## **Study 2 Decision Ladder Construction Semi-Structured Interview Questions**

### **Part 1: Decision Ladder Construction**

- How did you use the information from the alert entry to decide (Log/Block/Escalate)? (We will review the decision-making process for at least two alerts for each action category.) For a specific alert (probe by # alert code), what information influenced your decision to choose a particular action (Log/Block/Escalate)?
- When deciding whether to log, block, or escalate, what made you choose one action over the others?
- In general, how does the information provided by the interface influence your decision-making for most alerts?
- Can you describe a situation where you were uncertain whether to log, block, or escalate an alert? What factors influence your final decision most?
- Is there anything about the alert triage process that you find particularly challenging? If so, what specifically makes it challenging, and why?
- Are there times when you feel the information (alert-related and runbook information) is insufficient? If so, could you describe those situations?
- When making decisions, how often do you refer to the runbook rules and criteria? Which rules or information are the most difficult to recall when making decisions?
- Which pieces of information or features in the interface do you find most helpful in speeding up or facilitating your decision-making process?

### **Part 2: General Questions**

- Is there anything else we haven't discussed that you'd like to mention? Do you have any further questions or thoughts?

### **Part 3: Follow-up Questions**

In addition to the above questions, we reviewed selected alert actions with each participant and asked follow-up questions about their assessment process. We focused particularly on cases involving

longer processing times, decision changes, or instances where participants revisited and modified their decisions after reviewing other alerts.

## Study 2 Alerts List

Note that the alert in rows with a white background represent defined pattern alerts, while rows with a light grey background indicate ambiguous pattern alerts.

**Table 15.** Alerts Selected for Study 3.

Alert ID	Attack Type	Severity Level	Firewall Logs	Anomaly Score	IDS/IPS
#1	DDoS	Medium	Blocked Connection, IP Range Mismatch	Low	DDoS Attempt Detected
#2	Intrusion	Low	Blocked Connection, Unauthorized IP	High	Suspicious Packet Detected
#3	Malware	Medium	Blocked Connection, Unauthorized IP Medium	High	Anomaly Detected
#4	DDoS	Medium	Blocked Connection, Unauthorized IP	High	DDoS Attempt Detected
#5	DDoS	Low	Blocked Connection, IP Range Mismatch	High	DDoS Attempt Detected
#6	Malware	Medium	Blocked Connection, IP Range Mismatch	Low	Suspicious Packet Detected
#7	Intrusion	Low	Accepted Connection, Protocol Match	High	Intrusion Detected
#8	Intrusion	High	Blocked Connection, Unauthorized IP	Low	Intrusion Detected
#9	Malware	Low	Blocked Connection, Unauthorized IP	High	Malware Signature Match
#10	DDoS	Medium	Accepted Connection, Protocol Match	High	Intrusion Detected
#11	DDoS	Low	Blocked Connection, IP Range Mismatch	High	Anomaly Detected
#12	Intrusion	High	Blocked Connection, Unauthorized IP	High	Intrusion Detected
#13	Intrusion	Low	Accepted Connection	Low	Malware Signature Match
#14	DDoS	Medium	Blocked Connection, IP Range Mismatch	Medium	DDoS Attempt Detected
#15	Malware	Low	Blocked Connection, High Anomaly	High	Malware Signature Match
#16	Malware	High	Blocked Connection, Unauthorized IP	Low	Malware Signature Match
#17	Malware	Medium	Blocked Connection, Unauthorized IP	Low	Malware Signature Match

## Study 2 Supplementary Materials: Runbook (FAQ)

The following content is displayed on the alert triage interface ([link](#)) and can be accessed at any time from the sidebar under "Runbook (FAQ)".

The content is organized into four sections:

### Part 1: Response Action Guidance

This part explains what happens to an alert after each response action is taken.

Response Action Guidance			
Action	When to use	Criteria	Result
Log	The alert is suspicious but does not require immediate action. Select "Log" to keep track of the alert for additional signs of malicious behavior or further evidence.	<ul style="list-style-type: none"> <li>• <b>Non-critical Malware Signature Matches</b></li> <li>• <b>Low-severity alerts or routine incidents</b> without clear malicious indicators.</li> <li>• <b>Routine Protocol Matches:</b> Aligning with expected traffic patterns or previously identified benign anomalies</li> </ul>	The alert is logged and recorded in the system, but no further action is required.
Block	The alert indicates <b>clear malicious behavior</b> (e.g., malware, DDoS, suspicious traffic).	<ul style="list-style-type: none"> <li>• <b>High Anomaly Levels or Unauthorized Access Attempts:</b> Involving repeated violations or unauthorized IPs.</li> <li>• <b>Traffic Associated with Malicious Activity :</b> Including known DDoS patterns, repeated unauthorized attempts, or alerts for malware, intrusion attempts.</li> </ul>	Stop the traffic or quarantine a compromised system. The system or network traffic is blocked, isolating the threat and preventing further damage
Escalate	An alert is <b>too complex or high-risk</b> for you to resolve. Escalate when the alert is with <b>significant threat indicators, high severity, or unconfirmed behavior</b> merit further investigation.	<ul style="list-style-type: none"> <li>• <b>High-severity intrusion or malware</b></li> <li>• <b>Malicious Patterns or Signatures:</b> Especially those flagged in IDS/IPS with repeated Indicators of Compromise (IoCs) or known advanced threats.</li> </ul>	The alert is passed on to more experienced team members for advanced investigation, ensuring serious threats are properly handled.

### Part 2: Criteria.

This part outlines the clear patterns of key indicators and the recommended action decisions.

Criteria Box					
Attack Type	Severity Level	Firewall Logs	Anomaly Score	IDS/IPS Alert	Recommended Action
DDoS/Intrusion	Low	Accepted Connection, Protocol Match	*less than 30		Log
Malware	Low			Malware Signature Match	Log
Malware	Medium	Unauthorized Access Attempt	*less than 30		Block
DDoS	Low / Medium	IP range mismatch		Anomaly Detected	Block
Malware / Intrusion	High		more than 50	Malware Signature Match	Escalate
Malware	Medium	High Anomaly	more than 50	Malicious Pattern Match	Escalate

**Table 16.** Recommendations from Runbook.

Attack Type	Severity Level	Anomaly Score	Firewall Logs	IDS/IPS Alert	Recommended Action
DDoS	Low	<=30	Accepted Connections / Protocol Match		Log
DDoS	Medium		IP Range Mismatch	Anomaly Detected	Block
Intrusion	Low	<=30	Accepted Connections / Protocol Match		Log
Intrusion	High	>50			Escalate
Malware	Low			Signature Match	Log
Malware	Medium	>50	High Anomaly	Signature Match	Escalate
Malware	High	>50		Pattern Match	Escalate

## Part 2: Entry Classification and Related Information

The part describes the meaning of each column in the alert entries, including explanations and common values.

**Entry Classification and Related Information**

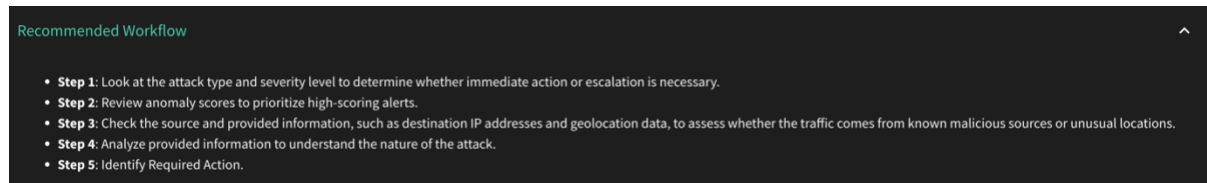
Alert Entry Title	Description & Common Values
<b>Alert Type</b>	<ul style="list-style-type: none"> <li><b>DDoS</b> - Denial of Service attempts detected by abnormal traffic volumes</li> <li><b>Malware</b> - Indicators of Compromise (IoCs) indicating malicious activity or known malware patterns.</li> <li><b>Intrusion</b> - Unauthorized access attempts or known attack signatures.</li> <li><b>Anomalies</b> - Out-of-pattern network behaviors, potentially a precursor to an attack.</li> </ul>
<b>Severity Level</b>	<ul style="list-style-type: none"> <li><b>Low</b> - Minimal impact with low risk, often involving non-urgent issues or informational events.</li> <li><b>Medium</b> - Moderate risk, usually pointing to potential security concerns that may require attention but are not immediately critical.</li> <li><b>High</b> - Critical threats with high impact and urgency, likely active threat.</li> </ul>
<b>Firewall Logs</b>	<p>Firewall logs provide additional context about whether traffic has been blocked, allowed, or redirected. They can help assess if an ongoing attack was already mitigated or if additional actions are needed.</p> <ul style="list-style-type: none"> <li><b>Protocol Match</b> - Traffic aligns with expected communication protocol.</li> <li><b>IP Range Mismatch</b> - Source IP falls outside allowed IP range.</li> <li><b>Accepted Connection</b> - Connection permitted by firewall rules.</li> <li><b>Blocked Connection</b> - Connection denied due to firewall restrictions.</li> <li><b>High Anomaly</b> - Unusual activity indicating potential threat.</li> <li><b>Unauthorized IP</b> - Access attempt from an unapproved IP address.</li> </ul>
<b>Attack Signature</b>	<p>A specific pattern or sequence of behaviors in network traffic or system activity that indicates a known cyberattack or malicious activity.</p> <ul style="list-style-type: none"> <li><b>Known Pattern A</b> - Often associated with a specific, targeted vulnerability or a well-known exploit (e.g., targeting a particular application, version, or protocol). Alerts for Pattern A suggest a precise, repeatable attack method, where blocking or patching the specific vulnerability may mitigate further incidents.</li> <li><b>Known Pattern B</b> - Generally indicates a broader attack strategy that may adapt or morph (e.g., reconnaissance or brute-force methods targeting a range of systems or ports). Pattern B may not target one specific weakness but aims to find any exploitable entry.</li> </ul>

<b>Anomaly Score</b>	<p>This score quantifies how unusual or suspicious an activity is. Higher scores indicate a greater deviation from normal behavior, suggesting a higher likelihood of threat.</p> <p>Score Range:</p> <ul style="list-style-type: none"> <li>• <b>0-30</b>: Low anomaly, likely benign.</li> <li>• <b>31-70</b>: Moderate anomaly, warrants further review.</li> <li>• <b>71-100</b>: High anomaly, potentially malicious, needs immediate attention.</li> </ul>
<b>IDS/IPS Alert</b>	<p>This is a notification from an Intrusion Detection/Prevention System indicating potentially harmful network activity.</p> <ul style="list-style-type: none"> <li>• <b>Malware Signature Match</b> - Identified known malware by its unique signature.</li> <li>• <b>Malicious Pattern Match</b> - Detected activity aligning with known malicious behavior.</li> <li>• <b>Intrusion Detected</b> - Unapproved access attempt or intrusion into the network.</li> <li>• <b>Suspicious Packet Detected</b> - Unusual packet flagged that may indicate risk.</li> <li>• <b>DDoS Attempt Detected</b> - Distributed Denial-of-Service attack suspected.</li> <li>• <b>Anomaly Detected</b> - Activity deviating significantly from typical patterns.</li> </ul>
<b>Timestamp</b>	Exact time the alert was triggered.
<b>Source IP Address</b>	IP address initiating the traffic.
<b>Destination IP Address</b>	IP address receiving the traffic.
<b>Source Port</b>	Port number used by the source device.
<b>Destination Port</b>	Port number on the destination device.
<b>Protocol</b>	<p>Communication protocols are standardized rules and formats that enable devices to exchange data efficiently and reliably over networks. They define how data is formatted, transmitted, and received, ensuring that different systems (like computers, IoT devices, or vehicles) can understand and respond to each other's messages.</p> <ul style="list-style-type: none"> <li>• <b>TCP</b> (Transmission Control Protocol) - A connection-oriented protocol that ensures reliable, ordered data delivery between devices.</li> <li>• <b>UDP</b> (User Datagram Protocol) - A connectionless protocol that sends data without establishing a connection, making it faster but less reliable than TCP. UDP is used in scenarios where speed is prioritized over reliability.</li> <li>• <b>ICMP</b> (Internet Control Message Protocol) - Primarily used for diagnostics and network management, ICMP helps report errors and send control messages. ICMP traffic generally consists of control packets rather than actual user data.</li> </ul>

<b>Packet Length:</b>	Size of the packet in bytes.
<b>Packet Type</b>	<p>Type/category of packet. Control packets and Data packets serve different purposes:</p> <ul style="list-style-type: none"> <li>• <b>Control</b> - Control packets can reveal unauthorized connection attempts or disruptions.</li> <li>• <b>Data</b> - Data packets help detect data leaks, malware, and unusual data transfers.</li> </ul>
<b>Traffic Type</b>	<p>Specify the purpose or content of the communication (web, file transfer, domain resolution).</p> <ul style="list-style-type: none"> <li>• <b>HTTP/HTTPS</b> - HTTP and HTTPS are used for web-based services within vehicles, such as infotainment, navigation, over-the-air (OTA) updates, and cloud connectivity for remote diagnostics. Alerts may indicate unauthorized access attempts, data exfiltration, phishing, or man-in-the-middle (MITM) attacks targeting vehicle applications or services.</li> <li>• <b>FTP</b> (File Transfer Protocol) - FTP-related alerts could point to unauthorized data uploads/downloads, malware payload transfers, or vulnerabilities due to the lack of encryption in standard FTP.</li> <li>• <b>DNS</b> - DNS resolves domain names to IP addresses, enabling vehicles to connect to cloud services, navigation servers, and IoT devices. DNS alerts may highlight DNS spoofing, DNS tunneling (used for data exfiltration), or phishing attempts where malicious domains mimic legitimate services.</li> </ul>
<b>Payload Data</b>	Content within the packet.
<b>Malware Indicators</b>	<p>Signs or evidence suggesting the presence of malicious software on a network or device</p> <ul style="list-style-type: none"> <li>• <b>IoC (Indicator of Compromise) Detected</b> - it means specific evidence (like a known malicious IP, domain, or file hash) suggests compromise but doesn't confirm active malware presence.</li> </ul>
<b>User Information</b>	Associated user or account details.
<b>Device Information</b>	Identifying info of the involved device.
<b>Network Segment</b>	Specific network area where traffic originated.
<b>Geo-location Data</b>	Physical location of the source IP.

## Part 4: Recommended Workflow

This part provides a suggested five-step brief process for handling an alert entry through to the final decision.

A dark-themed panel titled "Recommended Workflow" with a small upward-pointing arrow in the top right corner. It contains a bulleted list of five steps for handling an alert entry.

- **Step 1:** Look at the attack type and severity level to determine whether immediate action or escalation is necessary.
- **Step 2:** Review anomaly scores to prioritize high-scoring alerts.
- **Step 3:** Check the source and provided information, such as destination IP addresses and geolocation data, to assess whether the traffic comes from known malicious sources or unusual locations.
- **Step 4:** Analyze provided information to understand the nature of the attack.
- **Step 5:** Identify Required Action.

Participants are free to navigate to this page whenever they need to consult the provided information while completing the experiment task.

## Study 2 Participant List

This table provides a summary of the 12 participants in Group 1 of Study 3.

**Table 17.** Study 3 Participants (Group 1) List.

Participant ID	Work Experience	Cybersecurity Knowledge
P01	Engaged with cybersecurity-related content without a primary role in the domain (10+ years)	Has taken cybersecurity-related education or training without holding a formal degree or specified certification
P02	Novice	Limited
P03	Novice	Limited
P04	Novice	Limited
P05	Engaged with cybersecurity-related content without a primary role in the domain (Years of work experience not reported)	Has taken cybersecurity-related education or training without holding a formal degree or specified certification
P06	Novice	Limited
P07	SOC analyst, 5+ years of experience	Reported holding a cybersecurity-related certifications without specifying the name
P08	SOC analyst, 2+ years of experience	Reported holding a cybersecurity-related certifications without specifying the name
P09	Cybersecurity specialist, 2 years of experience	Reported holding a cybersecurity-related certifications without specifying the name
P10	Cybersecurity audit specialist (electric vehicle sector, 3+ years)	Reported holding a cybersecurity-related certifications without specifying the name
P11	Cybersecurity analyst (electric vehicle sector, 8+ years)	CISSP, CEH
P12	Cybersecurity QA tester (electric vehicle industry, 5+ years)	Reported holding a cybersecurity-related certifications without specifying the name

## Study 2 Production Rules Descriptions

Table 18. Production Rules Descriptions

Stage #	Rules	Rule Description
0 ACTIVATION: alert	0_retrieve_alert_position	<i>IF</i> the goal buffer state is detect_new_alert; <i>THEN</i> update the goal buffer to find_new_alert, retrieve a chunk from declarative memory for stored alert position information
	0b_find_new_alert	<i>IF</i> the goal buffer state is find_new_alert, and the retrieval buffer has successfully retrieved a chunk from declarative memory of an unattended alert position. <i>THEN</i> check the <u>visual location</u> buffer to add a new visual location chunk to represent the alert, update the goal buffer state to checking_alert_text_pos
1 OBSERVE: key indicator	1_check_alert_text	<i>IF</i> the goal buffer state is checking_alert_text_pos, and the visual location buffer is full as successfully encoded a visual location; <i>THEN</i> update the goal buffer state to <u>attending alert text</u>
	1b_move_attention_to_text	<i>IF</i> the goal buffer state is attending_alert_text, and the visual location buffer contains the alert text's position (_visuallocation) and The visual buffer is free; <i>THEN</i> issue a move_attention command to the visual buffer to shift focus to the alert text and update the goal buffer state to processing_alert_text.
	1c_wait_for_alert_text_encoding	<i>IF</i> the goal buffer state is processing_alert_text, the visual location buffer contains the alert text's position (_visuallocation), and the visual buffer holds the alert text value <i>THEN</i> store the alert text (val) in the imaginal buffer as attended_info, clear the visual buffer (~visual>), update the visual location buffer to the next position, and update the goal buffer state to move_attention_to_sev.
	2_move_attention_to_sev	<i>IF</i> the goal buffer state is move_attention_to_sev, and the imaginal buffer contains the alert text, and the visual location buffer holds a new valid position (_visuallocation), and the visual buffer is free <i>THEN</i> add a new visual command (_visual) to move attention to the severity level location stored in the visual location buffer and update the goal buffer state to processing_severity.
	2b_wait_for_severity_encoding	<i>IF</i> the goal buffer state is processing_severity, the visual location buffer holds a valid position (_visuallocation), and the visual buffer has this perceived value (val). <i>THEN</i> store the severity level (val) in the imaginal buffer (attended_info), clear the visual buffer (~visual), and update the visual location buffer with a new focus position, and update the goal buffer state to move_attention_to_anomaly,
	3_move_attention_to_anomaly	<i>IF</i> the goal buffer state is move_attention_to_anomaly and the imaginal buffer contains attended information, meanwhile the visual location buffer holds a valid position (_visuallocation) and the visual buffer is free <i>THEN</i> update the goal buffer state to processing_anomaly, add a new visual command (_visual) to move attention to the stored visual location of anomaly score.
	3b_wait_for_anomaly_encoding	<i>IF</i> the goal buffer state is processing_anomaly, and the visual location buffer contains a valid position (_visuallocation) and the visual buffer contains an encoded value of the displayed anomaly score (val). <i>THEN</i> update the goal buffer state to encode_anomaly_level, and store the anomaly score (val) in the imaginal buffer under attended_info.
3	4_processing_anomaly_level	<i>IF</i> the goal buffer state is encode_anomaly_level and the imaginal buffer contains the anomaly score, and retrieval buffer has the encoded anomaly level.

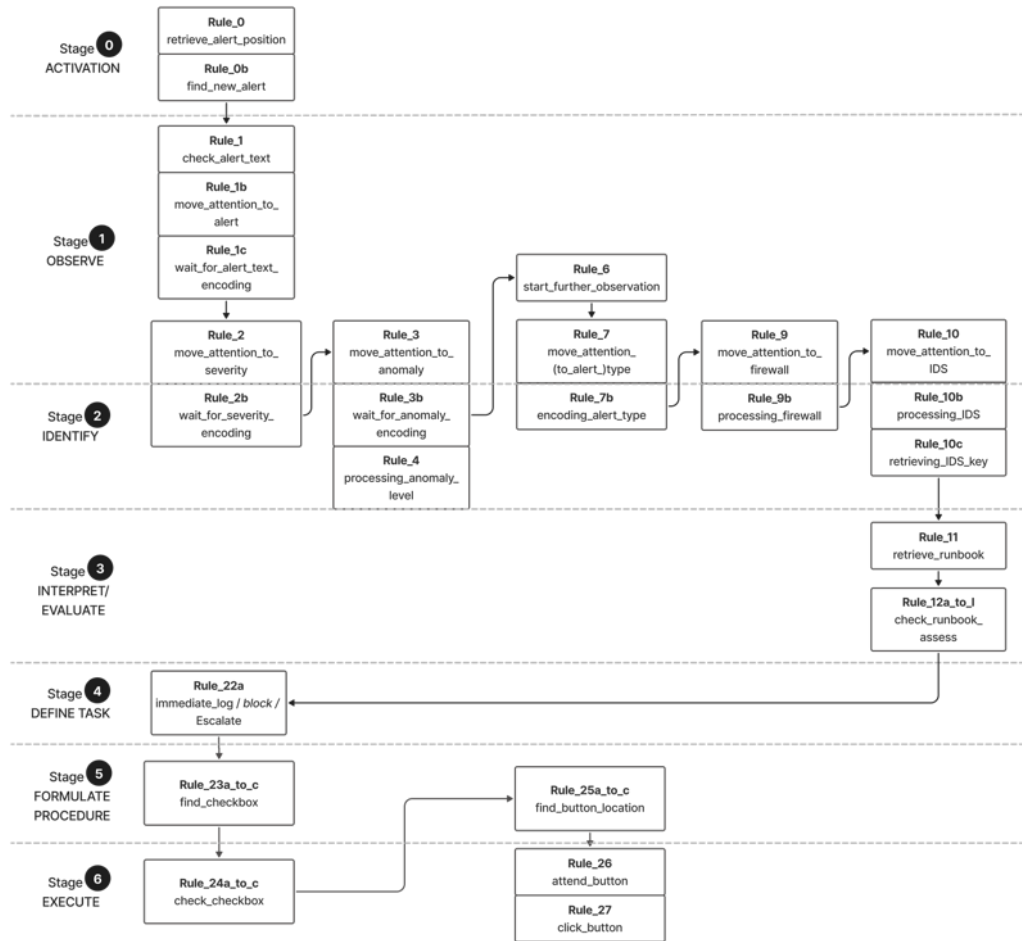
INTERPRET/ EVALUATE:  Anomaly Level		<b>THEN</b> update the goal buffer state to confirm_anomaly_level
2  IDENTIFY:  Alert Pattern	4b_init_assessment	<b>IF</b> the goal buffer state is encode_anomaly_level and the imaginal buffer contains attended information. <b>THEN</b> update the goal buffer state to initial_decision for the initial round of indicators patterns.
3  INTERPRET/ EVALUATE:  Alert Priority	5a_to_c_init_risk_a ssessment	<b>IF</b> the goal buffer state is initial_decision, and now the imaginal buffer contains attended information, including alert_text, anomaly_score and severity_level; <b>THEN</b> retrieve a chunk where status_decision is High/Medium/Low as the interpretation the indicators matching any patterns and update the goal buffer state to formulate_list_Escalate/Block/Log accordingly.
	5d_init_risk_assess ment_fall_back	<b>IF</b> the goal buffer state is initial_decision (and none of the above condition are met), <b>THEN</b> update the goal buffer state to confirm_anomaly_level
1  OBSERVE:  Threat & IOC & more cyber intelligent data	6_start_further_obs ervation	<b>IF</b> the goal buffer state is confirm_anomaly_level and the retrieval buffer contains the encoded values, and imaginal buffer attended information, <b>THEN</b> update the goal buffer state to move_attention_to_type, and , update the visual location buffer with a new position of next indicator-attack_type, and update the goal buffer state to move_attention_to_type.
	7_move_attention_t ype	<b>IF</b> the goal buffer state is move_attention_to_type, the visual location buffer has a valid location and the visual buffer is free; <b>THEN</b> update the goal buffer state to processing_type and add a new visual command chunk (_visual) to move attention to the stored attack type visual location.
	7b_encoding_attack _type	<b>IF</b> the goal buffer state is processing_type, and the visual location buffer contains a valid location, and the visual buffer holds a detected value (val). <b>THEN</b> store (val) as the attack_type in the imaginal buffer, and clear the visual buffer, update the visual location buffer with a new position of next indicator- firewall logs, and update the goal buffer state to move_attention_to_firewall.
3  INTERPRET/ EVALUATE:  Alert Priority	7c_encoding_attack _type_assessment	<b>IF</b> the goal buffer state is processing_type, and the visual location buffer contains a valid location, and the visual buffer holds a detected value (val). <b>THEN</b> store (val) as the attack_type in the imaginal buffer, clear the visual buffer, and update the goal buffer state to initial_decision_type.
	8a_to_c_init_risk_a ssessment_ (attack_type)	<b>IF</b> the goal buffer state is initial_decision_type, and now the imaginal buffer contains attended information, including alert_text, anomaly_score, severity_level and attack_type; <b>THEN</b> retrieve a chunk where status_decision is High/Medium/Low as the interpretation the indicators matching any patterns and update the goal buffer state to formulate_list_Escalate/Block/Log accordingly.
	8d_init_risk_assess ment_fall_back	<b>IF</b> the goal buffer state is initial_decision_type; <b>THEN</b> update the goal buffer to move_attention_to_firewall.  (The utility is set to -1 so that if the conditions of the above rule are not met, this rule will serve as a fallback, firing when the alert pattern does not match the retrieved criteria;)
1  OBSERVE:	9_move_attention_t o_firewall	<b>IF</b> the goal buffer state is move_attention_to_firewall, and the imaginal buffer contains attended the displayed information, the visual location buffer holds a valid location and at the same time the visual buffer is free;

Threat & IOC & more cyber intelligent data		<b>THEN</b> update the goal buffer state to processing_firewall, add a new visual command chunk to move attention to the stored location of firewall logs.
2 IDENTIFY: Indicators with prominent cues highlight the alert pattern	9b_processing_firewall	<b>IF</b> the goal buffer state is processing_firewall, the visual location buffer contains a valid position and the visual buffer has a stored value (val). <b>THEN</b> update the goal buffer state to move_attention_to_IDS, store the retrieved firewall logs value (val) in the imaginal buffer, clear the visual buffer, and update the visual location buffer to focus on a new position.
1 OBSERVE: Threat & IOC & more cyber intelligent data	10_move_attention_to_IDS	<b>IF</b> the goal buffer state is move_attention_to_IDS, the visual buffer is free, <b>THEN</b> update the goal buffer state to processing_IDS, and add a new visual command to move attention to the visual location of IDS/IPS.
2 IDENTIFY: Prominent Indicators cues highlight the alert pattern	10b_processing_IDS	<b>IF</b> the goal buffer state is processing_IDS, and the visual buffer contains a value, <b>THEN</b> update the goal buffer state to retrieving_IDS_key and store the IDS_IPS value in the imaginal buffer (attended_info).
	10c_retrieving_IDS_key	<b>IF</b> the goal buffer state is retrieving_IDS_key, and a retrieval request has successfully retrieved an ids_key which represents specific IDS/IPS pattern indicators, <b>THEN</b> update the goal buffer state to retrieve_runbook_round2.
3 INTERPRET/ EVALUATE: Alert Priority	11_retrieve_runbook  Alert Pattern, Missing & Conflicting Info	<b>IF</b> the goal buffer state is retrieve_runbook_round2, and both the retrieval buffer and the imaginal buffer are full, <b>THEN</b> update the goal buffer state to check_runbook_round2_assess. This means the model has successfully retrieved the necessary information and is now ready to assess it against the 2nd round of the runbook criteria.
	12a_to_1_check_runbook	<b>IF</b> the goal buffer state is check_runbook_assess, and the imaginal buffer contains attended information where alert indicators—including severity level, anomaly score, attack type, IDS/IPS alert, and firewall log keywords—match the predefined runbook criteria for categorizing alert patterns, <b>THEN</b> update the retrieved alert chunk by setting status decision to Low/Medium/High, and update the goal buffer state to define the appropriate action based on the identified pattern match.

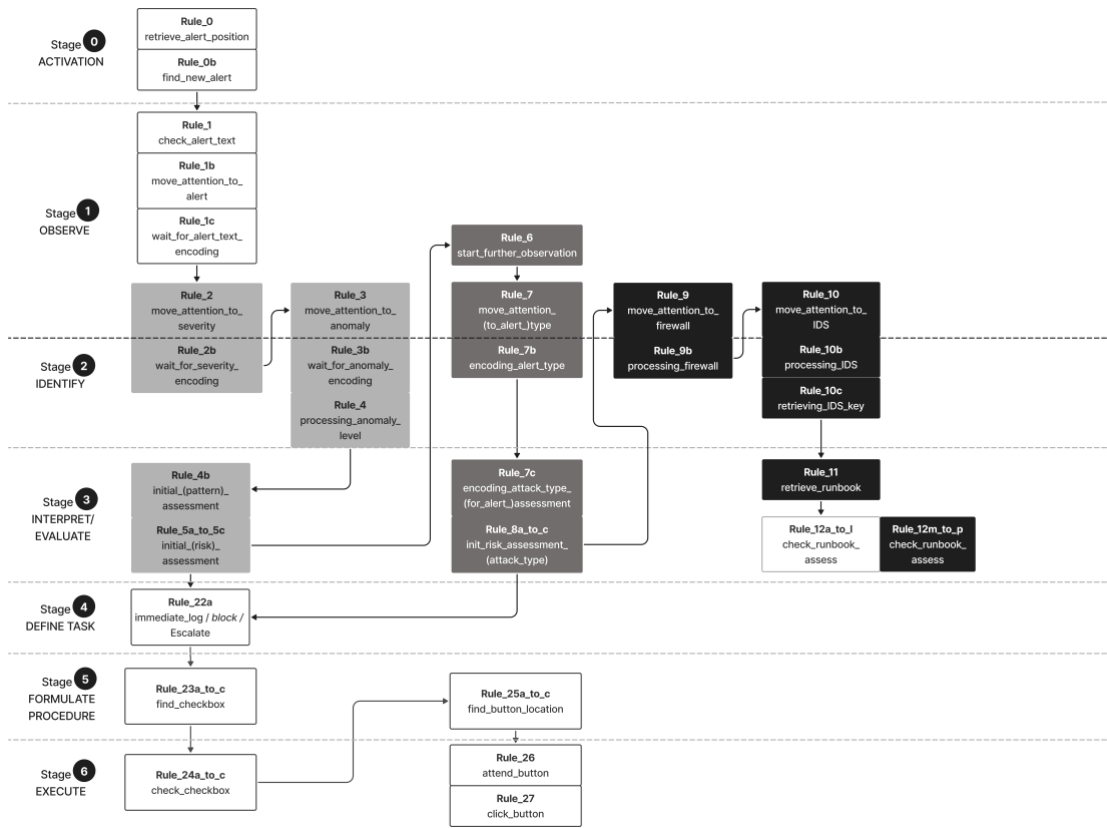
	12m_to_p_check_runbook	<p><b>IF</b> the goal buffer state is check_runbook_assess, and the imaginal buffer contains attended information where a subset of alert indicators combinations match part of the predefined runbook criteria for categorizing alert patterns,  <b>THEN</b> update the retrieved alert chunk by setting status decision to Low/Medium/High, and update the goal buffer state to define the appropriate action based on the identified pattern match</p> <p>(The utility is set to -1, reflecting the Streamlined Strategy, which has a lower likelihood of firing when the Pattern-Based Strategy is effective.)</p>
	21z_check_runbook_unknown	<p><b>IF</b> the goal buffer state is check_runbook_assess, and the imaginal buffer contains attended information where alert indicators—including severity level, anomaly score, attack type, IDS/IPS alert, and firewall log keywords—not match the predefined runbook criteria for categorizing alert patterns,  <b>THEN</b> update the retrieved alert chunk by setting status decision to unknown and update the goal buffer state to define the appropriate action as escalate.</p> <p>(The utility is set to -20 so that if the conditions of the above rule are not met, this rule will serve as a fallback, firing when the pattern does not match any runbook criteria);</p>
4	22a_immediate_log	<p><b>IF</b> the goal buffer state is the defined_action (Log/Block/Escalate), and the visual location buffer contains a stored location,  <b>THEN</b> update the goal buffer state to the action and set the visual focus to of the check box of the alert for further processing.</p>
DEFINE TASK:	22b_immediate_block	
Immediate Action	22c_immediate_escalate	
5	23a_find_checkbox_log	<p><b>IF</b> the goal buffer state is the action (Log/Block/Escalate), the visual location buffer contains a stored location of the check box, and the visual buffer is free (i.e., not currently attending to anything),  <b>THEN</b> add a visual command chunk to move attention to the stored visual location of check box and update the goal buffer state to check_checkbox for further processing.</p>
FORMULATE PROCEDURE:	23b_find_checkbox_block	
Locate Checkbox	23b_find_checkbox_escalate	
6	24a_check_checkbox_log	<p><b>IF</b> the goal buffer state is check_checkbox, the manual buffer is free (i.e., no ongoing manual actions),  <b>THEN</b> issue a manual command to press the key (simulating a checkbox selection) and update the goal buffer state search_list_button for defined action on which view to proceed with the next step for action button click.</p>
EXECUTE:	24b_check_checkbox_block	
Checkbox	24c_check_checkbox_escalate	
5	25a_find_log_list_button	<p><b>IF</b> the goal buffer state is search_list_button, the manual buffer and visual buffer are both free,  <b>THEN</b> update the visual location to (where the decision button is located) and update the goal buffer state to attend_decision_button to shift attention toward the decision button.</p>
FORMULATE PROCEDURE:	25b_find_block_list_button	
Locate Button	25c_find_escalate_list_button	

6 EXECUTE:  Button Click	26_attend_decision _button	<b>IF</b> the goal buffer state is attend_decision_button, the visual location buffer contains a position where the button position matches, <b>THEN</b> retrieve the alert chunk to update status_decision to Processed, add a visual command chunk to move attention to the visual location, update the goal buffer state to click_button to proceed the decision.
	27_click_button	<b>IF</b> the goal buffer state is click_button, the visual buffer is free (to confirm the visual attention shift is completed), and the manual buffer is free (i.e., ready to execute a manual action), <b>THEN</b> add a manual command chunk to press the key simulating the button click, update the goal buffer state to check_for_remaining_alerts to determine if another alerts need processing, clear the visual buffer to reset attention for the next task.

## Study 2 Mapping of Production Rules by Strategy Analysis and ConTA (Individual View of Different Strategies)



**Figure 48.** Production Map for only Pattern-based Strategy Rules.



**Figure 49.** Production Map for only Streamlined Strategy Rules (Different colors indicate distinct sets of alert features connected through a sequence of rules representing one of the strategies).

## Appendix C

### **A Follow-Up Study Extending the Integrated Framework: An exploration of GAI Models in Cognitive Modeling**

Since the start of this PhD work, cognitive modeling has also advanced rapidly in tandem with the development of AI (Malloy & Gonzalez, 2024; Niu et al., 2024). Recent frontier work (Collins et al., 2022; Joshi & Ustun, 2024; Malloy & Gonzalez, 2024; Niu et al., 2024) explores the integration of AI models with classical cognitive frameworks to estimate human-like decision-making. LLMs, as widely adopted GAI models, have drawn attention for their potential as cognitive models (Bandi et al., 2023; Binz & Schulz, 2023; Malloy & Gonzalez, 2024; Niu et al., 2024) and introducing novelty to classical models (Malloy & Gonzalez, 2024). One of the key advantages of GAI models is their demonstrated strength in exploratory capabilities through statistical pattern learning (Malloy & Gonzalez, 2024). More GAI models are believed to present significant potential in the sophisticated inductive capabilities to be integrated, but also being questioned about their limitations in covering decision-making opacity due to fundamental differences from human cognitive processes (Niu et al., 2024), training data quality requirements for task-specific sensitivity (Malloy & Gonzalez, 2024), human tendency to overinterpretation of model capabilities (Veres, 2022), and replicating subjective human biases (Zhao et al., 2022). Therefore, their performance may degrade in highly dynamic, real-world settings that fall outside the scope of their training. As such, researchers emphasize that relying on GAI models to replace cognitive modeling for simulating human decision-making remains unreliable at present (Malloy & Gonzalez, 2024).

Despite these concerns, integrating GAI has the potential to enhance traditional cognitive modeling in exploratory conditions (Malloy & Gonzalez, 2024). When faced with ambiguous pattern alerts, our human participants exhibited less structured decision-making paths, relying on knowledge-based reasoning and heuristics beyond rule-based modelling. In such cases, GAI models may be capable of handling these knowledge-based reasoning and heuristics and predicting decision behaviors (Nguyen et al., 2024). Building on this, we aim to examine whether large language models (LLMs), as a widely used form of GAI, can accurately estimate human decisions with ambiguous alerts triage and whether the reasoning processes offer exploratory solutions that extend beyond traditional cognitive models.

## Follow-up Study Objectives

This follow-up study aimed to: (1) evaluate whether the LLM can accurately predict alert triaging actions in handling ambiguous alerts and potentially outperform the integrated human cognitive model in defined-pattern alerts, and (2) compare the reasoning performance of the LLM and the integrated model.

## Experiment Settings

To assess performance differences between models, the analysis focuses on three dimensions: (1) the alignment between LLM-predicted decisions and our cognitive model's predictions of human responses to defined-pattern alerts; (2) the alignment between LLM-predicted decisions and actual human decisions for ambiguous pattern alerts; and (3) a comparison of the reasoning processes between LLM and human, focusing on key features and supporting evidence in decision-making.

This follow-up study used GPT-4-turbo (version: gpt-4-1106-preview) as a representative LLM to analyze the alert triage decision-making task of SOC Tier-1 analysts. ChatGPT-4 was selected as it is one of the most powerful and widely accessible LLMs and demonstrated strong general-purpose performance across various tasks and research works, including reasoning, summarization, and decision support (H. Liu et al., 2023).

Specifically, the model was prompted to evaluate each alert and triage as a tier-1 analyst. We used two structured prompts (Z. Li et al., 2025) to guide ChatGPT-4 in triaging decisions and reasoning (Z. Li et al., 2025) (All alert variables and the values were included in the prompts to enable both decision generation and reasoning analysis):

- **Introduction Prompt:** Please take on the role of a Cyber Security Operations Center Tier-1 Analyst and prepare to triage the provided alert list. Your task is to explain how the features of each alert contribute to the triage decision. Each alert includes various feature variables that should be considered in your analysis to determine the appropriate action, [INTRODUCE THE ALERT FEATURE NAMES AND DESCRIPTIONS, and THE RUNBOOK GUIDELINES].
- **Instruction Prompt:** For this alert profile [ALERT PROFILE], assess how each feature contributes to your triage decision. Your analysis should include no more than two key

identifiers; explanations of how the identifier(s) and other features of the alert support the actions.

For human participants, the key deciding and supporting features were identified through interviews with Group 1 (see below for the specific question; complete interview questions are provided in the Appendix B). The first two features mentioned for each alert were designated as the key deciding features to ensure comparability. Any additional features were categorized as supporting features.

- How did you use the information from the alert entry to decide (Log/Block/Escalate)? For a specific alert (probe by # alert code), what information influenced your decision to choose a particular action (Log/Block/Escalate)?

## Results

### Evaluating Model-Human Decision-Making Alignment

The comparison of majority decisions for each alert (respectively made by human participants, the integrated cognitive model), and decisions made by ChatGPT-4 (GPT-4-turbo) is summarized in Table 19, along with their F1 scores and accuracies. This includes results for defined pattern alerts and ambiguous alerts (comparing the human decision majority with ChatGPT-4 decisions).

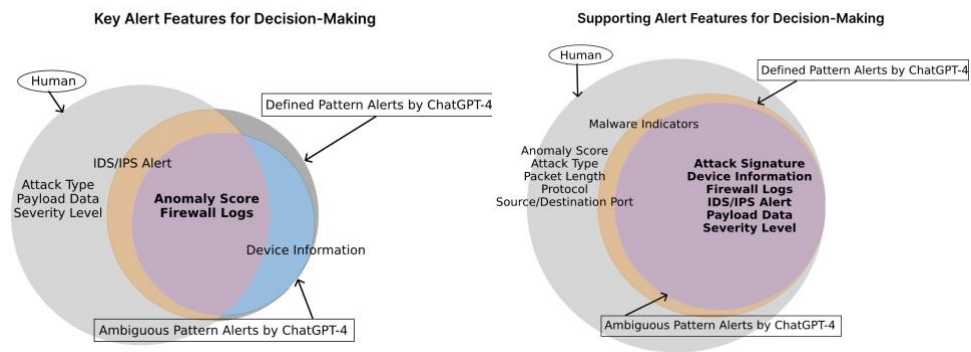
**Table 19.** The comparison of majority decisions among human, integration of CWA & ACT-R and ChatGPT-4 results.

Defined Pattern Alerts Decisions Modelling		
	Accuracy	F1 (Macro)
CWA+ACT-R vs. Human Decisions	0.8	0.79
ChatGPT-4 vs. Human Decisions	0.3	0.23
Ambiguous Pattern Alerts Decisions Modelling		
ChatGPT-4 vs. Human Decisions	0.43	0.43

Model accuracy was assessed by comparing each model’s decisions to the majority judgment made by human participants for each alert; F1 scores were calculated to evaluate how effectively each model predicted the same decisions as the human majority. Our integrated cognitive model demonstrated higher accuracy and F1 scores on defined-pattern alerts compared to ChatGPT-4, whereas ChatGPT-4 showed improved predictive performance on ambiguous-pattern alerts.

## Features used for Decision-Making

The comparison of alert profile features used for decision-making was conducted across three groups (human's selected features for all alerts, GPT-4's selected features for defined-pattern alerts, and GPT-4's selected features for ambiguous-pattern alerts) and two dimensions: key deciding features and supporting evidence features (see Figure 50). To check whether ChatGPT-4 processes ambiguous pattern alerts differently from defined patterns, we separated its identified features across the defined and ambiguous conditions.



**Figure 50.** Features Selection (key deciding features (Left) vs. supporting evidence features (Right)) between Human Participants, and ChatGPT-4.

As shown in the Figure 50, ChatGPT-4's selected features for ambiguous pattern alerts appear to be a subset of those used for defined pattern alerts, both in terms of key decision-making and supporting features. This suggests limited differentiation in feature processing between the two alert types, which may be partially attributed to the smaller number of ambiguous pattern alerts in the experiment. Moreover, we found that all supporting features identified by ChatGPT-4 were also reported by human participants (See Figure 50 (Right)), implying that human decision-makers may consider a broader range of supporting features. In terms of key deciding features, while there was a main overlap between humans' and ChatGPT-4's (See Figure 50 (Left)), ChatGPT-4 selects device-related information. In contrast, human participants relied on a broader array of features, which ChatGPT-4 ignored.

## **Discussions**

Returning to the purposes of the follow-up study, we evaluated the accuracy of the LLM model (ChatGPT-4 as a representative) in predicting Tier-1 analysts' decisions of alert triage. We compared it with our integrated model of CWA and ACT-R. The comparison was based on both models predicted decisions' alignment with the majority of human triage decisions for each alert. Results showed that for defined pattern alerts, the human cognition model provided more accurate estimations.

For the second objective, we examined the reasoning behind predicted decisions by ChatGPT-4. ChatGPT-4 provided structured justifications based on key deciding and supporting features, guided by the structured prompts. However, there was minor divergence between ChatGPT-4's key features used and humans', with human reasoning showing more diversity in feature consideration than ChatGPT-4.

### **Comparison of Models' Alignment with Human Decision-Making**

ChatGPT-4's performance shows much lower alignment with human triage decisions compared to our cognitive model for defined-pattern alerts. While its alignment improves slightly for ambiguous-pattern alerts, it still does not fully match the consistency of human decisions. The better performance of our cognitive model is not unexpected, as it was built directly on this specific, limited set of alerts and is overfitted to this task.

We also considered additional factors that may have contributed to ChatGPT-4's suboptimal performance on ambiguous-pattern alerts. From prior research has shown that general models like ChatGPT-4 often fail to adapt effectively to domain-specific workflows without proper fine-tuning, tailored prompting, or integration with up-to-date, domain-specific knowledge sources (X. Chen et al., 2024; Yang et al., 2023). Another possibility is that general LLMs are typically trained on expert-level data, which may not reflect the performance of novice users (R. E. Wang et al., 2024), just as our human participants. The underrepresentation of novice-level data is a critical limitation of LLMs in capturing adaptive learning mechanisms (J. Huang et al., 2024). In contrast, classic cognitive models such as CWA and ACT-R explicitly account for novice learning paths (Hassall et al., 2010; Lebière et al., 2019), as identifying bottlenecks and tracing errors is essential for supporting novice learning and guiding iterative, learning-driven design and development (Tulis et al., 2016).

### **Comparison of Selected Features in Decision-Making Processes**

The comparison of feature selections between humans and ChatGPT-4 reveals some differences in the reasoning.

One noticeable difference between human and ChatGPT-4 feature selection is that human participants treat 'Severity Level' as a key deciding factor for alert triage. In contrast, ChatGPT includes it only as a supporting feature. There are two potential explanations for this. First, the 'Severity Level' is visually emphasized in the interface through color coding and clearly labeled categories. Many participants reported that their attention was initially drawn to this feature due to its perceptual salience. Humans process visually salient cues efficiently, whereas the model receives only text-based input and thus does not prioritize visually salient information. Second, the 'Severity Level' is a synthesized indicator derived from tool-based assessments of historical data and recommended solutions, but it may not always be accurate. Yet novice users tend to over-trust this information (Passi & Vorvoreanu, 2022), treating it as a reliable indicator. Conversely, 'device information' is identified by the model as a key feature in the alert triage task, while humans treat it only as a supporting factor. This difference may also be due to the interface design, as device information is hidden within the expandable (accordion) section of the alert display, making it less immediately accessible to users. Additionally, the length and complexity of the device-related text may discourage human participants from using it quickly during decision-making. Both feature deviations between humans and ChatGPT-4 highlight how interface presentation can significantly influence the features humans rely on in decision-making.

We also find that human participants used a more diverse and broader set of supporting features in their decision-making compared to the model (See Figure 50 (Left)). ChatGPT-4 appears to narrow its focus to a comparatively smaller set of supporting features, while human participants tend to explore a broader range to support the decisions. We primarily attribute the model's more focused and structured feature selection to its training on high-quality, curated data. In contrast, the reasoning of our human participants may be less consistent in quality. We also doubt that the narrow selection of supporting features by ChatGPT-4 could also be attributed to the overgeneralization (Ralethe & Buys, 2022) and context saturation issues (where, with several rounds of prompts, large models tend to replicate the structures of previous answers) (Laban et al., 2025).

## **Limitations**

Two main limitations exist in this follow-up study. First, we used only two rounds of prompting to generate outputs from the LLM. Alternative prompting strategies, such as follow-up prompts, clarification questions, or prompts designed, were not explored and may have improved the model's output diversity and accuracy. Second, because this study was a post hoc analysis, we did not collect real-time feedback from human participants regarding the use of GAI models in the alert triage task, limiting the assessment of their perceived usefulness and accuracy in real human-AI settings.

## **Summary of the Follow-up Study Findings**

This follow-up study demonstrates that the classic cognitive models and LLMs exhibit different performance in predicting decision-making in the alert triage task.

The integrated traditional cognitive model performs better in estimating human decisions for clearly defined alerts than the LLM (e.g., ChatGPT-4). The better accuracy stems from both ACT-R's strength in well-structured perceptual, reasoning, and action cognitive architecture, and CWA's effective capture of the diversity and flexibility in real human decision-making of specific tasks.

On the other side, the main limitation of the classic integrated modelling approach is its challenge to simulate humans' nuanced handling of unstructured or ambiguous alerts within a rule-based symbolic structure. Whereas LLMs, specifically ChatGPT-4 used in this study, produced high-quality, well-structured reasoning for interpreting ambiguous pattern alerts, but its overall decision accuracy did not closely align with participants' responses to these ambiguous cases.

But its predictions tended to reflect only expert-level responses and a narrower reasoning structure, not covering the full range of decision-making by all participants. Notably, for ambiguous pattern alert decision-making, LLMs demonstrate improved alignment with human decisions than in clear pattern ones and more stable reasoning structures, narrowing the gap of our cognitive model in handling exploratory and unstructured reasoning processes. One reason for the better performance in the clear pattern alert of our traditional cognitive model is its task-specific construction, which may have overfitted to our participants within the defined task environment, in contrast to the more generalization of ChatGPT-4. Still, we are interested in any other factors may have also contributed to the underperformance of the LLM.

Given some observed deviations between ChatGPT-4's reasoning and human participants' decision-making processes, we further analyzed ChatGPT-4's reasoning to understand how alert entry features were processed in its decision-making. Two main factors may explain its deviation from the majority decision of our human participants. First, our comparison revealed discrepancies between the features selected by the LLM and those prioritized by human participants. This may stem from differences in the perceptual interface (visual interaction vs. text-based input), as human participants operate with visual stimulus cues on specific features such as color. LLMs rely solely on text-based context information and not prioritize perceptual salience as humans do (F. Han et al., 2025). Second, our participant pool includes a mix of novices and domain experts, whereas GPT-4 is primarily pre-trained on expert-level data. Other possibility of the low accuracy may be due to the model's contextual saturation (Laban et al., 2025) with our simple prompt design and the overgeneralization of large models (Ralethe & Buys, 2022). In conclusion, although ChatGPT-4's predictions may align more closely with expert-level decision-making, it demonstrates limited effectiveness in modeling the decision patterns of novice participants and lacks adaptation to the specific task domain.

### **Classic Cognitive Models vs. GAI Models**

Based on the brief follow-up comparison, the findings suggest future directions and potential enhancements for applying GAI models, informed by the strengths and insights of classical cognitive modeling.

**Enhancing Prompting for Reasoning:** While LLMs do not reason in the same way humans do, the step-by-step outputs often resemble human decision-making patterns by prompting strategies, such as CoT (Chain-of-Thought) (Wei et al., 2022), Tree-of-Thought (Yao, Yu, et al., 2023), and ReAct (Yao, Zhao, et al., 2023). These improved reasoning capabilities have been proposed not only to improve model accuracy but also to enhance model interpretability in human-AI collaborative settings, enabling humans to trace, monitor, and validate AI decisions. Selecting and ranking decision-relevant is critical for evaluating the faithfulness of the model's actual internal reasoning and causal structure (Bilal et al., 2025). As such, feature selection can be considered a foundational step toward aligning LLM reasoning patterns with the human-centric interpretability of prompt design.

Our follow-up study's feature selection comparisons indicate that ChatGPT-4's current feature selection tends to be narrower and more structured than our participants. We primarily attribute this to LLM's potential overgeneralization (Ralethe & Buys, 2022), underrepresentation of novice users (N.

Liu et al., 2023), and context saturation (Laban et al., 2025), and novice users' perceptual salient feature preferences and expert's thinking beyond the displayed information. These deviations illustrate the fundamental difference between human reasoning and LLMs: LLMs are constrained by the provided context, information, and pre-trained data, whereas humans are constrained by working memory, perceptual stimulus (ACT-R), knowledge and experience beyond the task scope (CWA). Given the distinct information processing approach, one future direction is to incorporate the CWA framework and ACT-R perceptual mechanisms into prompt design strategies, enhancing LLMs' ability to simulate human reasoning under task-specific constraints with improved task environments ecological validity, multimodal adaptively, and the explainability of model outputs.

**Error Tracing and Self-Refinement:** Human participants demonstrate a high level of self-reflection in our study, adjusting their final decisions based on an internal assessment of earlier data processing and learning (see Section 6.7.5). This ability to adapt decisions in response to new information is a core characteristic of human problem-solving (Amabile, 1983; Flower & Hayes, 1981). Iterative self-refinement involves generating an initial draft or decision and improving it through self-evaluation and revision (Madaan et al., 2023). Classic cognitive models are well-suited for capturing human performance limitations and errors within constrained work environments (CWA). ACT-R is also very effective in explaining and simulating human errors through its biologically grounded sub-symbolic mechanisms, including by memory decay and activation competition. Human errors or cognitive bottlenecks can be traced and explained using these validated cognitive modeling approaches, both internally (ACT-R) and externally (CWA). Beyond that, humans usually self-reflect on these errors and become adaptive through reasoning and learning mechanisms revealed by classic cognitive (Malloy & Gonzalez, 2024).

In contrast, the errors of large models are often less interpretable and more difficult to trace due to their inherently non-human-centric, statistically driven reasoning (Malloy & Gonzalez, 2024). Since LLMs are primarily trained on well-structured and expert-level data, the reasoning patterns of novice users may be underrepresented. Consequently, the error tracing and self-reflection inherent in the novice cognitive process are lacking in current LLMs. Besides, as current LLMs are pre-trained but remain static during inference, lacking real-time, self-adaptive learning mechanisms, unless an external adaptation framework is used (Jovanovic, 2024), adaptive learning in human cognitive models is also missing in these GAI models. This suggests a potential contribution from our traditional modeling work: human cognitive models could help inform error-correction and adaptive

learning mechanisms by capturing human error patterns, particularly among novice populations, for AI models.

**Human-Simulated Agent Development:** Furthermore, recent research on agentic interfaces and human-centered collaborative agents (e.g., Agashe et al., 2024; Mozannar et al., 2025) highlights the potential of our traditional cognitive models as essential tools for structuring step-by-step workflows that enable interpretable decision-making in task-specific human-simulated agents. Humans are not naturally adapted to verbal thinking and tend to perform worse when solving tasks verbally (Lombrozo, 2020). As for humans, key information is often encoded in visual and motor representations, but these are likely inaccessible to most current language models (R. Liu et al., 2025). This may somehow explain the underperformance of the LLM in our brief follow-up study. However, with ongoing advances in adapting multimodal AI models, agentic models are increasingly requiring the integration of various classical cognitive frameworks that are better suited to simulate human-like behaviors under a richer information representation context. This is also where our integrated model may extend its value as a robust human cognitive model featuring multimodal modules and domain specificity, that can inform the development of multimodal agentic systems within structured workflows.

The above discussions are based on our brief follow-up comparison between the integrated cognitive model and ChatGPT-4. Overall, as (R. Liu et al., 2025) say about research comparing human and GAI models, the comparison and discussions are not intended to validate GAI's potential to replace traditional human cognitive models. Rather, the value is in improving GAI models' performance by drawing on the distinctive insights from human cognitive models, such as error tracing, adaptive learning, feature selection, and prompt design optimization. The comparison analysis of the GAI model's application in simulating human performance can thus offer not only insight into how humans adopt and collaborate with AI, but also how AI models can learn from human cognitive models to become more robust, interpretable, and effective.