Formal Analysis of Vulnerabilities in Mixed-Reality Systems

Timothy E. Wang[1], Isaac Amundson[2], Junaid Babar[2] and Peggy Wu[1]

***Abstract*—With the proliferation of mixed-reality (MR) systems in aerospace and defense, there is increased potential for adversarial exploitation of system vulnerabilities and/or properties in the human cognitive process in order to reduce mission-effectiveness. This paper presents our preliminary work on the Modeling and Analysis Toolkit for Realizable Intrinsic Cognitive Security (MATRICS), a *formal methods*-based approach to provide a mathematically rigorous design and verification framework for protecting MR systems and operators in mission-critical applications from cognitive attacks. We describe our approach and present initial results, including formal models of the human operator, MR device, and mission environment, and apply existing formal methods tools to prove the holistic *cognitive security* of MR systems.**

I. Introduction

Mixed-reality (MR) and augmented-reality (AR) systems are becoming increasingly prevalent for aerospace and defense applications. For such mission- and safety-critical applications (e.g., military helmet-mounted displays (HMDs) [1] and the F-35 joint strike fighter HMD [2]), the required assurance may be considerably more complex than consumer-level applications. Traditional design-time assurance of mission- and safety-critical systems does not explicitly include a model of the human operator in the analysis but rather relies on assumptions about the behavior of the operator (e.g., the pilot is expected to perform predetermined tasks within a certain time period in the event of a specific contingency). It is foreseeable that in the context of military operations, potential malicious attacks from adversaries could be directed to exploit certain *cognitive vulnerabilities* of the human operator [3]. In addition to cyber attacks on devices, *cognitive attacks* executed in either the real or digital world could degrade or influence the performance of the human-machine system. Potential cognitive vulnerabilities in MR-human systems and the need for strong guarantees (i.e., high-assurance of mission success) motivates the development of formal methods [4] and formal cognitive behavioral models for a comprehensive security assessment of this emerging class of systems. There are challenges in validation of formal human cognitive models. For example, internal cognitive processes need large amount of experiments to validate and logistics constraints prohibits such experiments. To better align to the available, non-intrusive, and no-contact state-of-art measurement instrumentations, we are developing task performance models rather than very complex cognitive architectural models.

In this paper, we provide preliminary results on using formal methods to specify, analyze and design a HMD with provable guarantees of mission goals under cognitive attacks. We focus on visual perception attacks (i.e., cognitive attacks that degrade the performance of the operator doing a visual perception task such as detection of symbology on the MR/AR device display). The intended application of our effort is to develop formal models of the human operator that are useful for conducting a security assessment of the MR-operator system for a particular mission. The mission centers around a checkpoint for identifying adversary targets (i.e., vehicles, persons, etc.). The operator's task is to scan a surveillance area with the assistance of digital automated target recognition (ATR) overlayed on the HMD display to identify all valid targets. We present some background on MR/AR systems, cognitive vulnerabilities, formal methods for human-machine systems, and the formal methods framework used in this work in Sec. II. In Sec. III, we present our preliminary formal model of the HMD-operator system and the analysis results. Finally in Sec. IV, we discuss future work including planned refinements of the model.

II. Background and Reasoning Framework

A. Background and Related Work

Over the past thirty years, mixed-reality systems have rapidly evolved from proof-of-concept novelties to utilization in high-assurance domains such as aerospace [2] and healthcare [5]. This growth can be attributed to multiple factors, including advancements in computer vision, graphical processing and display technology, hardware miniaturization and inexpensive high-fidelity sensors [6]. The ability to

DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited. This work was funded by DARPA contract HR0011-24-9-0439. The views, opinions and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

[1]RTX Technology Research Center, {first.last}@rtx.com
[2]Collins Aerospace, {first.last}@collins.com

effectively process human and environmental input in combination with digital elements is also crucial for creating true MR experiences, necessitating significant contributions from the cognitive science and human factors domains.

The cognitive science literature includes studies going back more than half a century on the human response to various stimuli [7], [8]. More recently, in part driven by the video game industry, numerous studies have been conducted on cyber-sickness onset in immersive environments [9], [10]. However, far fewer studies have been published that focus on malicious cognitive attacks on operators of mixed-reality systems.

Deceptive techniques for influencing user behavior have become commonplace across the internet and typically involve confusing interfaces that mislead the user into providing personal information they might not otherwise disclose. Users of mixed-reality systems are susceptible to similar manipulations, and some of these *dark patterns* have been identified for VR [11] and AR [12] environments. A detailed exploration of the perceptual manipulation attack space is provided in [13]. However, none of these studies consider applying formal analysis to prove protection guarantees on cognitive attack models.

B. Formal Methods

The Assume Guarantee Reasoning Environment (AGREE) [14] provides model checking for systems modeled in the Architecture Analysis and Design Language (AADL) [15] with behaviors specified using assume-guarantee contracts. AGREE's compositional reasoning framework attempts to prove properties about one architecture layer using properties of components and subcomponents of the underlying layer. The composition is performed in terms of assumptions and guarantees provided as contracts for each component, where assumptions describe the expectations the component has on the environment and guarantees describe bounds on the behavior of the component when the assumptions are valid. AGREE is translated into Lustre [16] and analyzed using JKind [17], a k-induction model checker that supports multiple back-end solvers.

Formal methods have been used in prior works to analyze human-machine systems, albeit rarely, if ever, in the context of cognitive security. A survey of formal methods particularly for human-automation interaction can be found in [18] and a broader survey of the topic including formal methods for human-autonomy interaction can be found in [19].

III. Formal Models and Verification

In this section, we describe the formal HMD-operator model expressed using assume-guarantee contracts in AGREE. The overall model captures the mission task of a guard wearing a HMD visually confirming targets identified by ATR at a checkpoint. The ATR is a bounding box digitally rendered around the target on the HMD display. The goal of the adversary is to prevent the guard from seeing the ATR by directing an intense light source at the guard.

A. Modeling Assumptions

We make the following assumptions in the current model. Some of these assumptions will be relaxed in future models.

1) There is only one target in the field-of-vision (FOV) of the HMD at any given time. This greatly simplifies the model as we do not wish to model operator attention at this time.
2) There is only one source of light attack and it is localized around the attacker.
3) The perception process in the human operator is well-approximated by a reactive synchronous [16] process with a discrete-time clock, which is sufficient for the design and analysis of MR/AR devices.
4) The operator and HMD components are on the same clock (i.e., synchronized with the same sample period). A sample period of $20ms$ is assumed for both components. This assumption is made to simplify the AGREE analysis.

B. Cognitive Foundations

Changes in brightness can occur at a rate higher than what a human eye can adjust to, which can affect one's ability to perceive objects, either real or virtual. In a high tempo mission, the inability for the human to correctly distinguish between different types of symbology can have a detrimental effect on mission outcome. An average duration (596.4 ± 68.7 ms, 1607.6 ± 86.1 ms) of pupil contraction and dilation for healthy human subjects from 11-70 are observed in [20]. For the modeling of the operator's perceptual behavior, the pupil dilation and contraction period is currently used as a proxy for a bound on the duration in which the operator's ability to perceive symbols on the HMD display is compromised. We consider a brief bright light attacks during low light conditions similar to the preconditions in [20].

C. Details of the Model

A hierarchical illustration of our AGREE/AADL assume-guarantee model of the HMD-operator system is shown in Figure 1. A tutorial on the AADL/AGREE

tool could be found in [21]. The top-level component, *Checkpoint*, is a model of the mission. This top-level component is further refined into a composition of a *HMD* component and an *Operator* component. The *Operator* component formally captures a cognitive behavior of the human operator performing the mission. The *HMD* component is further refined into a composition of a mitigation component (*HMD Filter*), designed for reducing or eliminating the effects of the attack, and the display component (*HMD Display*), which detects targets and marks them with symbology (i.e., bounding boxes).

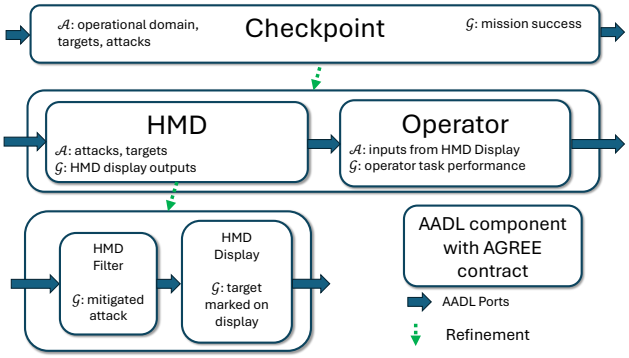

Fig. 1. Architecture of the overall HMD-operator model.

The assumptions and guarantees of components are expressed using AGREE and Lustre. The goals and objectives of the mission, i.e., the success criteria and operational domain of the mission (including the light-induced perception attack), are encoded in the assumptions, guarantees, and assertions of the AGREE contract. Assumptions are used to encode the operational domain and the set of possible attacks:

1) The ambient lighting condition is low, representing nighttime conditions.
2) The target is within the field of view.
3) At any given sample period, there is either an attack or there is no attack.

The guarantee expresses the criteria for mission success in the form of $P \rightarrow Q$, where the pre-condition P states that the target is persistently within the FOV for at least X ticks and the post-condition Q states that the mission task success is true for at least one tick.

For the *HMD* component, since it is comprised of the display and filter, its guarantees are a composition of the display and filter guarantees:

1) If the target is within the FOV for at least Y ticks, then the HMD display will mark it with a bounding box.
2) Otherwise, it will not.
3) Effective filter: after an intense light attack, the filtered light around the target converges to the ambient light within a short duration.
4) Ineffective filter: the filtered light around the target is equal to the perceived light around target.

In our model, there are two filter component implementations. The ineffective implementation does not filter out any light transients while the effective one always filter them out within a short duration. We specified an ineffective filter to illustrate that AGREE can uncover counterexamples in instances when MR device insufficiently protects against cognitive attacks. We also provide an effective filter to illustrate that AGREE could provide strong guarantees about a mission with a better designed MR device.

The operator's behavior in the presence of the light attack is captured by the following guarantees:

1) If a light attack of intensity α with duration between δ_1 and δ_2 ticks has occurred within the last Z ticks, then the operator does not detect the target.
2) Otherwise, the operator detects the target.

The time delay parameter Z, in which the operator's ability to detect the marked target on the display is degraded, is obtained by computing the $+2\sigma$ values of contraction and dilation duration. This time delay parameter along with other parameters of the guarantee (e.g., duration and intensity of the light attack) will be further refined using human subject experimentation. The formal AGREE contract of the *Operator* component is shown in Figure 2. The Lustre nodes *occurred* and *persistent* are used to capture certain time-bounded operators from past-time Linear Temporal Logic (LTL). The node *duration* is equivalent to *persistent* but for a range of durations. With this model, we can now formally analyze the top-level *Checkpoint* contract to either yield a possible cognitive attack in the form of an AGREE counterexample or guaranteed absence of such.

D. Formal Verification with AGREE

The model is publicly available [1]. We use AGREE's *Verify Monolithically* function to prove the following properties: consistency, compatibility and feasibility of the contracts, satisfaction of top-level mission contract by the composition of the lower-level contracts, and satisfaction of leaf node contracts by the implementations. For the MR model with an ineffective filter, the AGREE analysis (as expected) returns a counterexample representing a possible light attack

[1] https://github.com/loonwerks/MATRICS/tree/main/Models/Perception/contract_based/Perception

```
fun exceed(q1: real, q2: real, d: real) :
   bool = (q1 - q2) >= d ;
node occurred(cond : bool, t: int) returns (output : bool);
   var count : int;
   let
     output = if cond then true
       else if count <= t then (false -> pre(output)) else false;
     count = if not cond then (0 -> pre(count) + 1) else 0;
   tel;
node persistent(cond : bool, t: int) returns (output : bool);
   var count : int;
   let
     output = if count >= t then true else false;
     count = if cond then (0 -> pre(count) + 1) else 0;
   tel;
node duration(cond : bool, t1: int, t2: int) returns (output : bool);
   var count : int;
   let
     output = if (count <= t2 and count >= t1) then true else false;
     count = if cond then (0 -> pre(count) + 1) else 0;
   tel;
assume ambient_light_range "A: Ambient light between this range":
   ambient_light <= LOW_LIGHT and ambient_light >= 1.0;
guarantee target_detection "G: no attack means detection.":
   persistent(target_is_marked, 15) and
   not occurred( duration(
      exceed(filtered_light_around_target, ambient_light, 90.0),
      10, 36), 124)
   => detect_target;
guarantee no_target_detection
   "G: an attack means no detection for a period.":
   not persistent(target_is_marked, 15) or
   occurred( duration(
      exceed(filtered_light_around_target, ambient_light, 90.0),
      10, 36), 124)
   => not detect_target;
```

Fig. 2. The Operator component contract in AGREE and Lustre.

on the human operator. On the flip side, according to AGREE, the mission guarantee holds when the MR model with the effective filter is selected.

IV. Conclusion and Future Steps

We have presented our approach for the formal cognitive security analysis of mixed-reality systems. For future work, in addition to human subject experimentation to validate the operator component, we will create a more detailed implementation of the operator model and formally analyze the leaf operator component contract over this implementation. The detailed implementation will include various refinements to capture additional complexities involved in (1) the attack, (2) the lighting conditions (e.g., different colors instead of broad spectrum), (3) the external environment (e.g., occlusions and shadows), and (4) different perception modalities including auditory attacks. Furthermore, we will explore relaxation of some of the global modeling assumptions described in Section III-A (e.g., to include multiple targets and non-targets moving across the surveillance area in a dynamic fashion).

References

[1] M. M. Bayer, C. E. Rash, and J. H. Brindle, "Introduction to helmet-mounted displays," *Helmet-mounted displays: sensation, perception and cognition Issues*, pp. 47–108, 2009.

[2] Collins Aerospace, "F-35 Gen III Helmet Mounted Display System (HMDS)," *URL: https://prd-sc102-cdn.rtx.com/-/media/ca/f/f35/jsf-f35-datasheet.pdf*, 2025.

[3] A. Toet, "Optical countermeasures against human operators," in *Technologies for Optical Countermeasures XI; and High-Power Lasers 2014: Technology and Systems*, vol. 9251. SPIE, 2014, pp. 89–104.

[4] NASA Langley, "What is formal methods?" *URL: https://shemesh.larc.nasa.gov/fm/fm-what.html*, 2025.

[5] B. J. Park, S. J. Hunt, C. Martin, G. J. Nadolski, B. J. Wood, and T. P. Gade, "Augmented and mixed reality: Technologies for enhancing the future of ir," *Journal of vascular and interventional radiology (JVIR)*, vol. 31(7), p. 1074–1082, 2020.

[6] S. Vakkalanka, "A review on mixed reality operating systems : Current trends, challenges and prospects," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, pp. 53–61, 11 2024.

[7] E. Fehrer and D. Raab, "Reaction time to stimuli masked by metacontrast," *Journal of experimental psychology*, vol. 63, pp. 143–7, 02 1962.

[8] D. Simons and C. Chabris, "Gorillas in our midst: Sustained inattentional blindness for dynamic events," *Perception*, vol. 28, pp. 1059–74, 02 1999.

[9] R. Kirollos and W. Merchant, "Comparing cybersickness in virtual reality and mixed reality head-mounted displays," *Frontiers in Virtual Reality*, vol. 4, p. 1130864, 02 2023.

[10] N. Tian, P. Lopes, and R. Boulic, "A review of cybersickness in head-mounted displays: raising attention to individual susceptibility," *Virtual Reality*, vol. 26, pp. 1–33, 03 2022.

[11] W.-J. Tseng, E. Bonnail, M. Mcgill, M. Khamis, E. Lecolinet, S. Huron, and J. Gugenheimer, "The dark side of perceptual manipulations in virtual reality," *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 02 2022.

[12] X. Wang, L.-H. Lee, C. Bermejo, and P. Hui, "The dark side of augmented reality: Exploring manipulative designs in AR," *International Journal of Human-Computer Interaction*, vol. 40, 03 2023.

[13] K.-H. Cheng, J. F. Tian, T. Kohno, and F. Roesner, "Exploring user reactions and mental models towards perceptual manipulation attacks in mixed reality," in *USENIX Security Symposium*, 2023.

[14] M. W. Whalen, A. Gacek, D. Cofer, A. Murugesan, M. P. Heimdahl, and S. Rayadurgam, "Your "what" is my "how": Iteration and hierarchy in system design," *IEEE Software*, vol. 30, no. 2, pp. 54–60, 2013.

[15] P. H. Feiler and D. P. Gluch, *Model-Based Engineering with AADL: An Introduction to the SAE Architecture Analysis & Design Language*, 1st ed. Addison-Wesley Professional, 2012.

[16] N. Halbwachs, P. Caspi, P. Raymond, and D. Pilaud, "The synchronous data flow programming language lustre," *Proceedings of the IEEE*, vol. 79, no. 9, pp. 1305–1320, 1991.

[17] A. Gacek, J. Backes, M. Whalen, L. G. Wagner, and E. Ghassabani, "The jkind model checker," in *Computer Aided Verification - 30th International Conference, CAV 2018, Held as Part of the Federated Logic Conference, FloC 2018, Oxford, UK, July 14-17, 2018, Proceedings, Part II*, ser. Lecture Notes in Computer Science, H. Chockler and G. Weissenbacher, Eds., vol. 10982. Springer, 2018, pp. 20–27. [Online]. Available: https://doi.org/10.1007/978-3-319-96142-2_3

[18] M. L. Bolton, E. J. Bass, and R. I. Siminiceanu, "Using formal verification to evaluate human-automation interaction: A review," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 43, no. 3, pp. 488–503, 2013.

[19] T. E. Wang and A. Pinto, "Survey of human models for verification of human-machine systems," *arXiv preprint arXiv:2307.15082*, 2023.

[20] K. Tekin, M. A. Sekeroglu, H. Kiziltoprak, S. Doguizi, M. Inanc, and P. Yilmazbas, "Static and dynamic pupillometry data of healthy individuals," *Clinical and Experimental Optometry*, vol. 101, no. 5, pp. 659–665, 2018.

[21] "BriefCASE-tutorial," https://github.com/loonwerks/briefcase-tutorial, accessed: 2025-07-23.