Human Factors Design and Evaluation of Augmented Reality for Mass Casualty Incident Triage

Cassidy R. Nelson

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in the partial fulfillment of the requirements for the degree of:

Doctor of Philosophy
In
Industrial and Systems Engineering

Joseph L. Gabbard, Chair
Nathan K. Lau
James D. Ivory
Sebastian Oberdörfer

July 31, 2024
Blacksburg, Virginia

Keywords: Augmented Reality, User-Centered Design, Emergency Response, Mass Casualty Incidents, Triage, Implicit Bias

Human Factors Design and Evaluation of Augmented Reality for Mass Casualty Incident Triage

Cassidy R. Nelson

Abstract

Augmented reality (AR) is an emerging technology with immense potential for enhancing human-to-human interaction tasks, particularly in high-risk environments such as mass casualty incident (MCI) triage. However, developing practical and effective AR tools for this purpose necessitates a meticulous user-centered design (UCD) process, thoughtfully crafted and validated through iterative testing with first responders in increasingly contextually relevant simulations. In academic circles, the perceived complexity and time requirements of such a process might discourage its adoption within the constraints of traditional publishing cycles. This is likely due, in part, to a lack of representative applied UCD examples. This work addresses this challenge by presenting a scholarly UCD framework tailored specifically for MCI triage, which progresses seamlessly from controlled, context-free laboratory settings to virtual patient simulations and finally to realistic patient (actor) scenarios. Moreover, MCIs and triage are underserved areas, likely due to their high intensity and risk. This means developers need to 'get it right' as quickly as possible. UCD and evaluation alone are not an efficient means to developing for these complex and dangerous work domains. Thus, this research also delves into a cognitive work analysis, offering a comprehensive breakdown of the MCI triage domain and how those findings inform future AR supports. This analysis serves to fortify the foundation for future UCD endeavors in this critical space. Finally, it is imperative to recognize that MCI triage fundamentally involves human-to-human interaction supported by AR technology. Therefore, UCD efforts must encompass a diverse array of study stimuli and participants to ensure that the technology functions as intended across all demographic groups. It is established that racial bias exists in emergency room triage, creating worse outcomes for patients of color. Consequently, this study also investigates the potential impact of racial biases on MCI triage efficacy. This entire body of work has implications for UCD evaluation methodology, the development of future AR support tools, and the potential to catch racially biased negative performance before responders ever hit the field.

Human Factors Design and Evaluation of Augmented Reality for Mass Casualty Incident Triage

Cassidy R. Nelson

General Audience Abstract

Augmented reality (AR) is uniquely situated to make work within high-risk work environments, like mass casualty incident (MCI) response, safer and more effective. This is because AR augments the user's reality with context-relevant information, like by providing a temperature gauge for firefighters that is always in their visual field. Development of such AR tools for a sensitive arena like MCIs requires several rigorous steps before those tools can be deployed in the field. It is crucial to engage in a user-centered design (UCD) process in partnership with actual emergency responders so they can help us understand what help they need. We outline that UCD process in Chapter 2. Once we understand what responders say they need help with, we then need to evaluate those pinch points in the broader context of their work. This means that we evaluate how their job process creates the situation where responders need the kind of help they are asking for. Understanding this helps us create solutions that address the responder's needs while we minimize any new problems created with implementing a new tool into the job. What we learned from examining the work domain is described in Chapter 3. Once we have this firm foundational understanding of responder needs and work and we have designed a AR support tool, we need to evaluate that tool for effectiveness. It is too dangerous to put the AR tool straight into the field, so Chapter 4 explores how we can create simulations of an MCI scenario to study our AR support tool. Finally, after evaluating our AR tool within the scenarios and the scenarios themselves, we evaluate (in Chapter 5) other facets of the job that may be impacting MCI response. In our case, we explore how racial bias may be impacting patient care. It is important to study bias as it has implications for future MCI training and AR tool development. Perhaps future work can explore an AR tool that offsets bias-based performance, or a training that helps catch bias before responders ever get to the real field.

Table of Contents

Attribution

This work was done in collaboration with several colleagues. Most critically, this work was overseen, conceptually guided, and edited by Dr. Joseph L Gabbard; Kyle Tanous built the AR test-beds; The primary subject matter expert in emergency response is Dr. Jason Moats; Dr. Nathan K. Lau guided the cognitive work analysis; and Dr. Ranjana K. Mehta helped design and actualize data collection. The author of this dissertation guided the conceptualization, methodology, formal analysis, and writing of every phase of this work.

Funding Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2033592.

Acknowledgments

Before jumping into the dissertation, I want to express my heartfelt gratitude to those who have supported and encouraged me on my journey as an emerging scholar.

I am profoundly grateful to my advisor, Dr. Joseph L. Gabbard, whose guidance, wisdom, and unwavering support have been crucial to my growth as a researcher, scientist, and human being. I have deeply appreciated his mentorship in the ways of scholarship, collaboration, and how to enjoy the little things. Thank you for helping me explore.

I wish to thank my committee members, Dr. Nathan K. Lau, Dr. James D. Ivory, and Dr. Sebastian Oberdörfer, for their valuable suggestions, constructive criticism, and constant encouragement. Your expertise and perspectives have greatly enriched this work. I have enjoyed working with each of you immensely. To my entire committee: Thank you all for your approachability, guidance, and warmth. It has been a treat to schedule my defenses and learn from how you all balance levity and rigor.

Thank you to Dr. Ranjana K. Mehta and Dr. Jason Moats, who would have been on my committee if scheduling the required four professors was not already enough of a headache. Your mentorship and collaboration on this work has been instrumental. I look forward to continuing to work together to publish the work contained herein (though I hope not to be in Texas on any more black-flag weather days OR Wisconsin in the winter).

Thank you to my lab mates for the fun, for answering questions in the group chat late at night, and the shared celebrations of our accomplishments; Dr. Richard Greatbatch, Nayara de Oliveira Faria, Dr. Missie Smith (even though you graduated before I got here!), Kyle Tanous, Dr. Jared Van Dam, Dr. Eric Bloomquist, Obinna Ugwu, John Luksas, Alex Krasnwer, and Kelsey Quinn.

Thank you to the fantastic friends I have made in my wonderful time at Virginia Tech. We have literally traveled the world together for our academic pursuits (and for fun), and I could not have asked for better (or funnier) travel companions: Dr. Lee Lisle, Dr. Niall L. Williams, Matthew Gottsacker, Dr. Juan Diego Shiraishi, Alexander Giovanneli, and Zubin Choudhary. I am grateful for each of you.

Thank you to the mentors and colleagues who have been incredibly supportive of other facets of my academic development. I always look forward to seeing you at conferences and having nuanced conversations about our field; Dr. John Quarles, Dr. Tabitha Peck, and Dr. Rick Skarbez. I look forward to the next jam session.

Thank you to Hannah Parks, Nicole Conner, Akiko Nakata, Tina Russell, and any other faculty and staff that have facilitated my academic journey in complex, nuanced, and unseen (to me) ways. Thank you to every instructor and teaching assistant who taught me the foundations. Thank you to every undergraduate student who helped code data, transcribe interviews, and build test beds.

Thank you to my very patient family for listening to me drone on and on about my PhD work infinitum… and for their constant belief in me. Thank you to my best friend, Dr. Samantha Goodell, for her unwavering support in all facets of life… and for all the silliness.

Thank you to my heart, my little dog Lucy. She has been with me since my first semester of undergrad. She sits on my desk as I type this; her support and presence have been a constant warmth. She is grey in the face now—it has been a long 10 years together. I hope for many more. Thank you to my faithful laptop, that is duct taped together. We've written many papers and played many games together.

And last but certainly not least, an incandescent thank you to Wally Morris for his thoughtful support—be it late night proof reads, listening to me wax poetic about inane research rabbit holes, for all the reassurances, and for making sure I got plenty of sunlight and laughter in the midst of the most stressful waves. I love you.

Thank you to everybody who has ever loved me, supported me, or shown me warmth even in passing. I am grateful for every kindness.

To all my teachers, mentors, and anyone who has ever been kind to me.

1 Introduction: The Core of This Work

Collectively, this body of work addresses and advances the domain of AR for emergency response from multiple perspectives, each offering distinct methodological contributions. Contribution 1 outlines an exemplary academic user-centered design and evaluation process, and provides design considerations informed by emergency responder feedback. Contribution 2 explores the work domain of mass casualty incident triage and provides an ecological framing for future AR tool development in MCI triage. Contribution 3 explores the potential of different simulation strategies for measuring the AR and evaluates their effectiveness in tandem to the actual AR app (from Contribution 1) being tested. Finally, Contribution 4 explores the potential impacts of study stimuli on outcome measures in the simulations outlined in Contribution 3. More specifically:

Contribution 1: Applied User-Centered Design and Evaluation. This contribution details the development and implementation of an applied user-centered design and evaluation process for the design of an augmented reality (AR) triage tool suite for mass casualty incident triage. It emphasizes the importance of user engagement and feedback in refining AR tools for practical use. This chapter also describes AR use cases (how AR can enhance various aspects of emergency response), such as situational awareness, communication, and decision-making.

Contribution 2: Cognitive Work Analysis for Mass Casualty Incident Triage. In Contribution 2, a cognitive work analysis is presented, focusing on the ecological domain of mass casualty incident (MCI) triage. This analysis examines the cognitive processes, decision-making strategies, and information flow involved in triage scenarios. It provides insights into the challenges faced by emergency responders and offers design considerations for future AR systems aimed at optimizing triage efficiency and accuracy.

Contribution 3: Context-Intensive Simulation Development. Contribution 3 discusses the development and evaluation of two progressively context-intensive simulations designed to simulate realistic emergency scenarios. The simulations included one with virtual patients and one with ‘real’ patient actors, and provided the test bed for further evaluation of the Virtual Triage Tag described in contribution 1. We evaluate the simulations and compare them. Then we explore responder feedback provided via semi structured interviews. These simulations helped us capture a more comprehensive assessment of the VTT.

Contribution 4: Impact of Simulation Stimuli on Triage Accuracy. Contribution 4 explores the influence of simulation stimuli on the accuracy of emergency responders during triage tasks. By analyzing how different simulation factors (e.g., patient race/gender) affect decision-making and task performance, this contribution sheds light on potential factors that can change triage performance.

Together, these contributions provide a thorough exploration of AR technology's potential to enhance emergency response, grounded in user-centered design, cognitive analysis, and realistic simulation testing.

This holistic approach ensures that the AR tools developed are not only technologically advanced but also user-friendly and effective in high-pressure situations.

1.1 Why Mass Casualty Incidents

Mass casualty incidents (MCIs) often generate more patients than standardly available resources can manage with routine procedures, thus demanding "… exceptional emergency arrangements and extraordinary assistance" [1]. MCIs (e.g., tsunamis, nuclear meltdowns, bombs, etc.) are rarer than typical single-person events (heart attack, stroke, etc.) but more deadly due to the scale and severity of the event. This results in shortcomings in the response associated with poor management, leadership, planning, communication, and resource management [2]–[4].

Despite the wide range of potential MCIs, the public expects MCIs to be effectively managed [3]. This effective management entails a rapid response coordinated via triage to assess injured patients for the highest priority injuries to facilitate earlier care, greater medical resources, and evacuation priority [5], [6]. Emergency response triage during MCIs is cognitively demanding, physically exhausting, and dangerous due to the requisite demands in situation awareness, complex conformal relationships between resources and the unfolding MCI, and the precarious 'golden hour of trauma.' This 'golden hour' refers to a brief window of time where speedy, accurate, and effective treatment administration greatly impacts patient outcomes [7], connotating intense time pressure on responders to engage in a high-stakes event with often limited practice coordinating across counties, specialties, and mass triage itself (due to the rarer nature of MCIs). As if MCIs were not already complicated enough, we see even deeper and more concerning problems for patients with darker skin tones as emergency contexts are "… breeding ground[s] where clinical decisions are particularly influenced by implicit biases," [8].

Racial bias in healthcare is a pervasive phenomenon, with extensive evidence documenting anti-black bias in healthcare across specialties [8]–[13]. Substantial time has been allocated to studying this racial inequality between black and white patients [11], [14]. Some of these disparities stem from *implicit bias,* the unconscious, unintentional implicit associations between a group and a negative evaluation [15], leading to prejudice as a public health problem [16]. These biases impact quality [17], timeliness[13], and accuracy of care [18], [19] resulting in worse outcomes for black patients [8] even when accounting for socioeconomic status [9]. This trend continues into MCIs and worsens with the level of emergency [20]. Worded differently, the worse the emergency, the poorer the health outcomes for black patients. This is because implicit bias effects are exacerbated by the very same characteristics that make MCIs difficult in the first place, namely: time pressure, patient overload, and incomplete information [8], [21], [22]. Circling back to the 'golden hour' of trauma, triage is less accurate and slower for black patients in the emergency

room during MCIs [23], [24], meaning black patients are more likely to have the 'golden hour' expire before they receive appropriate care.

We posit that AR, specifically head-mounted and optical-see-through AR, is uniquely situated to offer the aforementioned "… exceptional emergency arrangements and extraordinary assistance" critical to MCIs [1]. AR has the unique ability to dynamically overlay context-sensitive virtual information upon the physical world while offering mobile, bare-handed, and head-up interactions [25]–[28] to facilitate more effective MCI triage for *all* patients. Head Mounted Display (HMD) AR within emergency response is an understudied arena, with most work focusing on mobile AR [25]. Thus, within emergency response, we seek to understand the use cases of AR, responder insights/needs, and design/evaluate AR tools for MCI response through user centered-design, usability evaluation, and a cognitive work analysis. Further, we explore the impacts of racial bias on MCI triage specifically to set a foundation for potential AR-based measurement and remediation.

1.2 Broader Implications and Context

User-centered design (UCD) and evaluation are critical to thoughtful development of human-computer interaction in novel technologies. To this end, researchers must understand the intended users and their needs, simulate usage with stakeholders in actual work tasks, and then find/fix problems discovered in user testing [29], [30] in a repeating cycle [31]. Unfortunately, in practice, this development cycle can appear too time-prohibitive on the front end, particularly in an academic research cycle that demands intensive publication output. Still, UCD is particularly critical in arenas where developers need to "get it right the first time" [31], e.g., high-risk work like is found with emergency response. As a result, we find emergency response to be a critically understudied area of AR application. **Part of what makes an iterative UCD development cycle time prohibitive is the 1) high-risk nature of MCIs demanding slower development, 2) the lack of scholarly, exemplary references applying a UCD for high-risk work, and 3) the need for intensive user-based simulation.** Given the nature of MCIs as complex and ever-evolving systems, we posit that usability evaluations alone are not sufficient (or at least, usability evaluations alone are a slower way to approach the problem). Thus, we offer a cognitive work analysis of the MCI triage domain to define the type of cognitive work afforded by responders, thus helping improve future AR (and other support technology) development.

From here, this need to 'get it right the first time' underscores the Cognitive Work Analysis (CWA) of the work domain. The UCD focuses on the end user and their tasks while the CWA focuses on the broader work domain. The difference here is that a CWA provides a fuller picture of the MCI triage space to help designers anticipate the needs of responders. A UCD cannot tell us how an intended solution will interact

with the broader work domain. Moreover, a UCD cannot give us a thorough understanding of underlying constraints that cannot be violated in the implementation of a new tool. **These complementary approaches offer a holistic picture of MCI triage** that, when taken together, offer design considerations from the work domain down to the individual responder.

Given the aforementioned high risk and complex nature of MCI triage, it is impossible and undesirable to test new AR tools in the actual field. But, it is also undesirable to go from a lab-based setting with representative tasks straight to the actual field, no matter how many lab studies are done. This is where different levels of simulations come in. We teamed up with Texas A&M's Extension Service at Disaster City to develop two simulations. One is a virtual patient simulation using AR patients, and the other is a 'real' patient actor simulation with real people pretending to be injured. Depending on the demands of the simulation, there is an implicit tension between more representative users/stimuli, meaningful UCD, and resource intensity. Thus, it is meaningful to understand the usefulness of potential 'virtual' simulations. **We evaluate these simulations as well as the VTT within them and describe strengths and weaknesses with each simulation, providing examples for future simulation development.** This work has implications for future simulation development, both for researchers trying to leverage simulation-based test beds, but also for future training development of emergency response simulations.

Finally, it is critical to note that in applied settings, AR is often not the 'primary task.' This means AR supports another primary/occupational task, for example, by providing firefighters with an external temperature gauge in the corner of their AR visual field to support appropriate fire mitigation strategies and safe egress. However, we must consider that a large subset of potential work tasks involve working with people, meaning that AR supports interaction with other people instead of machinery or infrastructure, as we find with emergency response. With an understanding of the work domain and responder self-reported needs, a critical next step is understanding how AR can facilitate the human-to-human interactions core to MCI triage. Central to intensive and effective user-based applications and simulations for human-to-human supported tasks, a diverse range of representative stakeholders are required. Unfortunately, scientific studies are often only conducted by and with the M-WEIRD (male, white, educated, industrialized, rich, democratic) representing only 3% of the global population. This trend extends to virtual reality research [32] and means that issues and nuances faced by other demographics are not represented in data or in study development, with implications for the representativeness of study stimuli. Understanding that AR can support human-to-human interactions, **it is critical to study the implications of diversity in the study stimuli itself.** This area is understudied and has sizable implications for UCD simulation/evaluation, training, and final realized AR integration. If bias can be captured in the simulations, this means that researchers need to

be thoughtful about study stimuli, and that training simulations could feasibly capture racial bias and mitigate it before responders ever get in the field.

1.3 Document Formatting and Study Design

This document comprises complementary chapters elucidating how we have 1) applied an iterative UCD strategy to develop an Augmented Reality Triage Tool Suite (ARTTS), 2) evaluated the tool suite's effectiveness within the MCI triage domain, as well as the merits of using virtual vs. 'real' patients (actors), and 3) used measures of performance defined in our second contribution (calculated with time to complete and accuracy) and exploratory analysis of the importance of diversity in study stimuli across virtual and 'real' contexts.

1.3.1 Chapter 2: User-Centered Design and Evaluation of ARTTS: an Augmented Reality Triage Tool Suite for Mass Casualty Incidents

As informed by emergency responders, this study outlines how AR is best situated to facilitate emergency response by determining use cases and the specific AR technology. Moreover, we offer an example UCD for the high-risk work of MCI triage response: the Augmented Reality Triage Tool Suite (ARTTS). Early iterations were evaluated qualitatively with a primary subject matter expert (SME), with a concluding analysis with ten responders for usability, workload, and engagement in a context-agnostic, controlled lab setting.

Objective 1: *Use a UCD development cycle to define and evaluate potential AR use cases for emergency response, expanding to design and assess an AR system to support emergency responders in MCI response.* In pursuit of this goal, we have the following research questions:

1. What are the use-cases for AR in emergency response?
2. What is the usability of the core Virtual Triage Tag (VTT) of ARTTS?
3. How does the VTT workload compare to workload associated with the current emergency response standard paper tag?

1.3.2 Chapter 3: A Cognitive Work Analysis on Mass Casualty Incident Triage: Lessons Learned and Design Considerations for Future AR Support Systems

Informed by the literature and our design process from chapter 2, we seek to dig deeper into the complex system of MCI triage. We agree with Vicente that "… A thorough work analysis must be conducted if computer-based systems are to effectively support human work in complex sociotechnical systems," [33] which MCI triage is. In so doing, we can offer deeper design considerations about the triage process and cognition within it, not just general design recommendations based on best usability as usability. A deeper

work domain understanding helps increase the effectiveness of future MCI AR designs, therefore setting each successive usability evaluation up for greater success.

Objective 2: Develop a deeper understanding of the MCI triage work domain and how the work domain interplays with design. From here, we ask the following research questions:

4. What defines the MCI triage domain from an ecological perspective?
5. How can AR integrate into MCI triage from an ecological approach?

1.3.3 Chapter 4: User-Centered and Situated Evaluation of an AR Virtual Triage Tag Comparing Virtual vs. Real Patients: Usability, Perceptions, and Performance

With a theoretical understanding of the MCI triage domain and an evaluation of the VTT's usability in a controlled, context-agnostic lab environment, we want to understand the usability and performance of this newly iterated VTT in a more context-intensive environment. More specifically, we want to ascertain the effectiveness of VTT's higher-order cognitive support elements in a medium-scale context (virtual environment) and a larger-scale context (live training demo) featuring a teammate while simultaneously comparing the virtual and real simulation for relative triage effectiveness. This study will evaluate usability, workload, cognitive workload, accuracy, completion time, and qualitative interviews with emergency responders.

Objective 3: *Investigate the VTT in a more context-intensive virtual and live training scenarios.* To do this, we ask:

6. How does the virtual patient simulation compare to the real patient actor simulation?
7. What is the usability of the VTT via quantitative measures of performance and qualitative reports from emergency responders?

1.3.4 Chapter 5: The Impacts of Bias on MCI Triage: Trends Across Virtual and Real Patients in Augmented Reality

Using the performance measures defined in objective 3, we now want to explore whether racial bias impacts triage performance. In tandem to the same experimental data captured in objective 3, we measured racial bias using the Implicit Association Task (IAT) and grouped via k-means clustering into "high" and "low" bias groups. We will further explicitly compare bias-based performance across virtual and real simulations.

Objective 4: *Explore whether we can capture racially biased performance in the simulations made in contribution 3.*

8. What are the impacts of racial bias on performance (defined in obj. 2) across virtual and real contexts?
9. Does racial bias impact the type of error classification?

Table 1: Planned Research Design

Objective	1. *Use an UCD development cycle to evaluate and define potential AR use cases for emergency response, expanding to design and assess an AR system to support emergency responders in MCI response.*	2. *Investigate the MCI triage sociotechnical system to understand how future AR can better support response.*	3. *Investigate the VTT in a more context-intensive virtual and live training scenario to elucidate both the VTT's effectiveness and to parse the usefulness (and therefore ultimate need) of human patient actor over virtual patient simulations.*	4. *Explore how racial bias manifests in the VTT and its implications for emergency response and future diversity in study design.*
Research Questions	*1. What are the use-cases for AR in emergency response?* *2. What is the usability of the core Virtual Triage Tag (VTT) of ARTTS according to first responders?* *3. How does the VTT compare to the current emergency response paper tag?*	*4. What defines the MCI triage domain from an ecological perspective?* *5. How can AR integrate into MCI triage from an ecological approach?*	*6. How does the virtual patient simulation compare to the real patient actor simulation?* *7. What is the usability of the VTT via quantitative measures of performance and qualitative reports from emergency responders?*	*8. What are the impacts of racial bias on performance (defined in obj. 2) across virtual and real contexts?* *9. Does racial bias impact the type of error classification?*
Independent Variables	*Classical paper tag vs. virtual triage tag*		*Virtual Patient Triage Task vs. Real patient (actor) triage task*	*Virtual Patient Triage Task vs. Real patient (actor) triage task*
Dependent Variables	• *Semi-structured interviews to define early prototypes* Prototype measured via: • *Usability measured with a System usability Scale (SUS)* • *Mental workload with a NASA-TLX* • *User Engagement via the User Engagement Scale (UES)*		• *Semi-structured interviews* • *Time to complete (TTC)* • *Accuracy (Number correct)*	• *Time to complete (TTC)* • *Accuracy (Number correct)* • *Implicit Association Task (IAT)* • *Error direction* • *P-score*
Research Task	*10 – Experience AR triage training. Experience measured test scenarios. Then semi-structured interviews on possible AR use cases in Emergency Response.* *11 – Complete a scripted list of representative tasks with the AR tag or paper tag. Do the same set again on the remaining condition.*		*Triage 5 virtual patients* *Triage 10 real patients*	*Triage 5 virtual patients* *Triage 10 real patients*
Experimental Design	*n = (10 + 11 = 21)* *Mixed subjects design, counterbalanced* IV1 – Classica paper tag vs. AR tag	Theoretical CWA guided by Vicente and Kilgore.	*n* = 13 *Mixed subjects design* IV1 – Virtual Triage Session vs. Live Triage Session	*n* = 1 *Mixed subjects design* IV1 – Virtual Triage Session vs. Live Triage Session

2 Study 1 – User-Centered Design and Evaluation of ARTTS: an Augmented Reality Triage Tool Suite for Mass Casualty Incidents

2.1 Introduction

The World Health Organization (WHO) defines mass casualty incidents (MCIs) as "… an event which generates more patients at one time than locally available resources can manage using routine procedures… requir[ing] exceptional emergency arrangements and additional or extraordinary assistance," [1]. A critical component in MCI response is triage, or the assessment of treatment and evacuation priority of injured persons [34], with poor coordination leading to increased mortality [35] via delayed triage [36]. Delaying triage is detrimental as there is a 'golden hour' of trauma where treatment efficacy greatly impacts outcomes [7], where effective triage processes are needed to increase survival chances for all injured parties by assuring highest priority injuries are assessed first [5], and resources are managed effectively. As augmented reality (AR) is a novel technology that affords mobile, bare-handed, head-up interaction with information superimposed atop the real, physical world [26], we posit AR can provide the additional, extraordinary assistance necessary to enhance the triage component of MCI response. To that end, we present ARTTS, *a head-worn Augmented Reality Triage Tool Suite* containing an initial sorting tool, virtual assessment tool, and virtual triage tag to assist emergency responders (ERs) in mass casualty incidents (described in more detail in Section 2.4).

To create ARTTS, our design team coordinated with an emergency response subject matter expert (SME, specifically hereafter referred to as our primary SME), wielding nearly 30 years of experience in not only emergency services but also emergency response training. Further, we leveraged additional subject-matter expert interviews from ten ERs after each had exposure to select AR experiences designed to illustrate relevant AR capabilities (e.g., annotating a scene to mark hazards, viewing head-up data, interacting with hand-based menus, etc.). We then identified several broad areas where AR could enhance emergency response and specifically improve triage processes. From these areas, we ascertained requirements and developed ARTTS. Then, we conducted a user evaluation of the triage tag's user interface and usability with eleven ERs.

Initial sorting acts as a simple series of screen prompts to kick-start triage activities by reminding novice ERs of the very first step of triage. These tasks are currently held only in ERs' memory, therefore relying on ERs' ability to recall process details and experiences/training while under duress. The virtual assessment tool (VAT) is more involved, housing the entire first wave triage decision tree that is also currently held only within a trained ER's memory. ARTTS' VAT provides a "click-through" presentation of the Sort, Assess, Lifesaving Interventions, and Treat-ment/Transport (SALT) framework [37]. *Virtual*

triage tags replace state-of-the-art physical paper tags currently employed by ERs, which are filled out with hand-written information by responders and attached to MCI patients to track recovery, treatment, evacuation, and triage efforts during MCIs. ARTTS' virtual triage tags denote personal patient information, injury type and severity, provided treatments, patient ID, patient triage order, and generally aim to overcome shortcomings of physical triage tags. For example, paper triage tags can become lost or illegibly damaged due to weather, feature a finite and small area in which to write, only allow unidirectional changes (worsening), [38] are static, and are difficult to see in the dark. Static refers to the paper tag's inability to automatically update themselves — they require constant manual res ponder input. Paper triage tags also do not provide a capability to prioritize patients within the same category [39] since the data contained on each tag cannot currently be aggregated or "rolled up" to a central server (for comparative analyses and summarized data views of the overall response), nor provide a physical space to manually denote additional prioritization. Finally, paper triage tags alone do not allow for advanced notice to hospitals. In our vision, ARTTS virtual triage tags will provide hospitals with advanced notice of patient injuries prior to their arrival, thus permitting enhanced, adaptive, specialized preparation.

This work contributes to the evolving body of knowledge related to using AR for emergency response by providing ARTTS, a list of general AR requirements and use cases for emergency response, specific AR-based triage requirements, a use case describing the application of a user-centered design and evaluation processes when integrating AR tools into existing occupational workflows, design principles for AR systems to support MCIs, and lessons learned from the evaluation that benefit other AR practitioners interested in fielding AR to non-technical user communities. We seek to answer:

1. How can we apply a user-centered design process to design for emergency response?

2. What are the use-cases for AR in emergency response?

3. How might these AR use-cases facilitate MCI triage specifically?

4. What is the usability of the core Virtual Triage Tag (VTT) of ARTTS in a lab setting, according to first responders?

5. How does the VTT compare to the current emergency response standard paper tag?

6. What design considerations can be gleaned from further ARTTS iteration?

In the remainder of this chapter, we first provide the limited amount of related work we found on using AR to support emergency response and specifically triage. Section 2.3 describes our user-centered design methods, development process, and the ER standards anchoring ARTTS. Section 2.4 illustrates the proposed ARTTS interfaces with additional discussion. In Section 2.5, we outline the results of our usability

evaluation conducted with actual ER personnel as well as design principles for AR systems. Lastly, section 2.6 delivers a conclusion encapsulating limitations and future work for ARTTS, as well as for AR in triage generally.

2.2 Related Work

As alluded to above, our review of the literature revealed opportunities to build upon the limited prior art, particularly in relation to AR support for initial sorting, SALT processes, and virtual triage tags. We did, however, find a few inspiring works of interest.

Fernández, Bernabe, & Rodriguez's [5] application leveraged Google Glass to provide decision support to first line responders through icon prompts for each step of the triage process. As a responder moves through the flowchart, a different icon glows, eventually leading to a category assignment. However, as the icons have no written prompts, novice responders may still struggle, particularly if the responders are not familiar with the symbology. This is compounded by the entire triage process, represented by icons, being always present. Sometimes, the process requires that responders skip steps, but this may be confusing for novice responders.

Vassell et. Al., [36], also leveraged Google Glass to capture triage tag identifiers (they used QR codes instead of bar codes) and generate an incident commander view of patient locations and severity. However, their QR code identifiers completely replaced the entire physical tag, not just the bar codes. While incident command information is enhanced and the system allows responders to select a QR marker printed in a color associated with the patient's triage category, all other tag information is lost to first-wave responders. Moreover, the limited field of view and monocular viewing limitations of google glass suggest there are opportunities to explore AR interfaces that leverage current state-of-the-art binocular displays with wider fields of view.

Nestler, Huber, & Klinker [11] propose a radio-frequency ID (RFID) enhancement to traditional triage tags, but this primarily aids incident commanders in assessing general ground status (number of patients and their rough locations) and does not directly aid first-line responders.

Killeen et. Al., [38], proposed an addition to the physical tag via a digital version displayed on personal digital assistants (PDAs). Scanning the physical tag barcode pulls up the digital version. The digital tag also provides decision support by leveraging the Simple Triage and Rapid Treatment (START) triage protocol [40]. However, PDAs require both the responder's hands to hold the device and the stylus needed to interact with the device. Further, START is no longer the standard of practice protocol for triage [37].

Mizumoto et. Al., [39] created and leveraged a bio-sensor to afford dynamic data collection for triage-tag information population that can be viewed with video-passthrough AR via a tablet. Again, the device requires one or both responders' hands and leverages the START protocol. Moreover, the triage tag was replaced with a drastically limited version containing only five information categories, all of them providing quantitative data (i.e., heart rate, respiration, etc.). This approach arguably results in a loss of all qualitative data and, in turn, provides much less information than the traditional paper tag affords.

Uchiyama et. Al., [41], designed a different type of AR marker leveraging blinking LEDs, affording a greater allowable distance between responders and patients while still allowing access to patient data tags. Again, the triage tag was completely replaced with a digital tag containing only quantitative data, which could be viewed as resulting in drastically diminished informational capacity in comparison to paper triage tags. Additionally, this system leverages a mobile device and thus requires at least one responder hand to hold and use the system.

Other work with AR and triage focuses on areas such as augmenting real-time data movement between incident commanders and responders [42], an AR training environment for MCIs in virtual hazards [43], providing advance status information en route to MCI sites [44] or to hospitals [7], or supporting more limited triage associated with profoundly austere conditions and non-medically trained 'responders' or lay-men [45].

2.3 User-Centered Methodology

In designing ARTTS, we first assembled a design team that featured several disciplines, including human factors engineering, usability/user experience engineering, computer science, visual arts, and emergency response. We employed a UX design cycle [46] (also outlined in figure 1) and embraced a human-centered design [47] philosophy for all ARTTS framework elements. In summary, we first sought to understand users and their work needs, then created candidate design solutions, and then iteratively prototyped and evaluated these solutions; at each step putting human needs, behavior, and capabilities first [47].

Our primary SME provided insight via interviews as well as reference materials for training in triage [48], and for existing identified areas for AR support in policing (a first-responder occupation with overlap to emergency response) [49]. From these resources and a literature review of related work, we extracted a categorized list of ways in which AR could support emergency response (see list in Section 4). Our primary SME then reviewed the categorized list and helped us prioritize AR tools for triage according to impact, feasibility, and likelihood of adoption. We were careful to consider where and how AR could complement existing triage processes instead of upending them as SMEs indicated technology acceptance as a consistent hurdle in emergency response. Our remining user-centered design activities for this phase

of work centered specifically on triage for MCIs. Wielding this adjusted list, we conducted a set of subsequent interviews with ten other SMEs asking about triage pinch points, how AR could alleviate them, and the requirements of any new system for adoption.

We then conducted a thematic content analysis using the qualitative transcript data and notes taken from the interviews with the ten additional SMEs. Consistently identified attributes were grouped and synthesized into work activity notes that we then organized into a work activity affinity diagram (WAAD). From the WAAD, we distilled user and flow data models [46] as well as personas in which to drive subsequent user interface design decisions. We leveraged the data models to capture system requirements for ARTTS.

We further anchored our requirements and emerging conceptual designs with established national emergency response standards [8] and work artifacts [50]. For example, ARTTS' conceptual foundation is heavily informed by the Model Uniform Core Criteria (MUCC) for Mass Casualty Incident Training [51], further ensuring that ARTTS could be integrated in and potentially even used for training in the future.

Wireframes (figure 2) helped quickly identify visual clutter and accidental emergent features for removal, as well as pinpoint areas for improvement for the transition from physical tag to virtual tag.

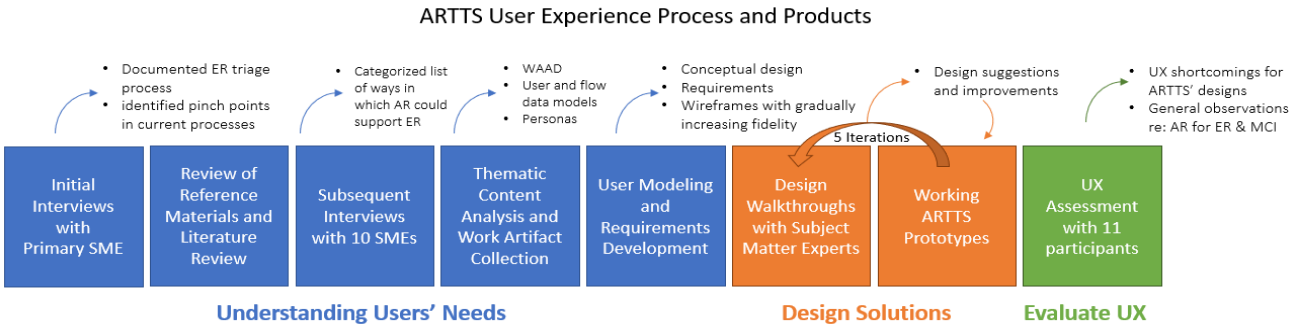

Figure 1: We employed a user-centered design and evaluation process to increase likelihood of ARTTS' adoption

Finally, ARTTS has come to prototype fruition several times, with expert insight for each iteration evaluated via design walkthroughs with our primary SME for changes. In sum, the version of ARTTS presented in this chapter underwent five design iterations.

2.3.1 AR Display Hardware

While mobile devices are smaller, cheaper, and more ubiquitous, SMEs identified it mission critical that ERs have their hands free to give medical interventions necessary for survival. In this vein, Coskun et. Al., [52] found that holding a device with one hand and using the other to interact with the device reduces device acceptance. Additionally, handheld tools can be easily lost as medical care requires the device to be pocketed or placed nearby and can also be dropped between patients.

MCI events are typically chaotic, and potentially dangerous, until ERs have established an evacuation plan and the threat has ceased, which sometimes does not occur until after evacuations have been completed. As such, we chose a binocular optical see-through AR display to support stereoscopic viewing of near field objects (specifically the user interfaces associated with ARTTS' tools) and assist in preserving situation awareness as replacing or diminishing reality was not identified as a requirement [53]. With these requirements in mind, we chose to leverage the HoloLens 2 as the medium for ARTTS current delivery, and provide a sample of AR display tradeoff considerations in Table 1.

Table 2: Sample of AR Display Tradeoff Considerations

Requirement	**Handheld AR**	**Headworn AR**
Hands-free to give medical interventions	One hand required to hold display and see information; second hand required to interact; potential issues with arm fatigue	Required for gesture-based interaction only
Aid in scene processing	Some integrated image and depth sensor	
Visualization of real-time data	Viewed indirectly via video pass-through AR	Viewed directly with via optical see-through AR
Facilitation of data collection relevant to search and also future training	Can leverage integrated GPS, eye tracking not available without additional hardware	Can leverage integrated eye tracking, GPS often missing
Facilitation of communication between response teams	May not perform well under loud conditions and/or limited privacy without use of additional hardware (which can be lost, require charging, etc.)	Can employ integrated microphone and speakers proximal to users' mouth and ears for better channel quality and privacy
Suitable for use in chaotic environments	Handheld size has advantages, but can be easily dropped and misplaced	Headword size has disadvantages, but is less likely to be misplaced
Provide situational awareness (SA) support	Must first alert users via vibration or tone before AR visual SA support can be provided	Is already on' head, such that visual notifications and alerts can be attended to instantly
Support hazard avoidance and general wayfinding	Requires looking through handheld as a lens onto the world, as well as spatial mapping between phone view and real-world view	Provides a natural first-person point-of-view perspective of integrated AR and real-world

2.4 Initial Design Results and Discussion

2.3.1 **ER Use Cases that can Benefit from AR**

By reviewing the literature and engaging emergency response SMEs we identified the following emergency response dependent needs/uses. Use cases identified in the supplementary materials and/or in our literature review are noted with citations, all others were derived from primary SME discussions and SME interviews.

Setting the stage (first waves of response. Establishing crisis extent, setting up evacuation points, finding patients)

- Facilitation of quick data collection and effort-free use
- Aids in scene processing
- Can implement a specific decision tree system for reporting

Communication

- Recording patient status
- Updating patient status
- Visualization and real-time editing of real-time data [54]
- Facilitation of interdisciplinary communication between specialists [54]
- Able to identify areas already searched [54]
- Facilitation of data collection relevant to search and also future training [54]
- Tracking of location, activity, and status of other responders [49]
- Integration of civilian data [55]
- Triage support [3], [7]
- Facilitation of communication between response teams across multiple disaster events [3]
- Survivor treatment support [3], [7]
- Streaming of patient information (video and audio) to be viewed by hospital responders [7]
- Advanced hospital notice

Decision making

- Evacuation progress summary
- Resource inventory
- Resource allocation tracking
- Jolting stunned responders back to triage
- Lessen cognitive load for novice responders in remembering triage process
- Visualize non-physically visible structures (pipes underground, etc.) [54]
- Lessening of cognitive load associated with personal information tracking [54]
- Support via third party with more complicated data that's been synthesized to provide guidance [54]
- Situational awareness support [54]
- Expert advice in real time from expert not physically responding [7]

Navigation/wayfinding

- Hazard avoidance
- Direct to points of interest [54]
- Further information about points of interest (type of point, distance, direction relative to user) [54]
- Map displays to support navigation [49]
- Thermal/Infrared [49]
- Advanced optics for zoom [49]

Threat detection

- Quick sharing of dangerous / damaged areas to fellow responders during response [54]
- Real-time intelligence about crimes/criminals in an area [49]
- Integration of substance sensors to notify of contamination [49]
- Friend identification worn by responders and picked up by helmet [49]

Training

- High training transfer ratio between device and actual use
- Quickly instituted
- Easy to train in-house (within a station. No third-party institutions needed)

2.4.1 General System and User Interface Requirements

Through work activity notes we identified three primary user types: novice responder, expert responder, and incident commander. Novice responders are ERs with limited experience, particularly in relation to MCI's. This limited experience renders novice responders more vulnerable to becoming stunned. Comparatively, expert responders define ERs with moderate to extensive experience, again particularly in relation to MCI's. This relation to MCIs is critical as the sheer scale of MCIs make these incidents profoundly different from typical, small patient pool response. Finally, incident commanders are responsible for higher-order aspects of emergency response including managing operations, resource allocation, and making decisions regarding risk/reward for search and rescue of hazardous environments. Incident commanders are not on the front lines and are instead typically located in a communication hub overseeing all incoming streams of data to inform decision making and using the afforded communication channels to communicate those decisions quickly to front line responders.

Current pinch points are mostly associated with evacuation status and consistent communication regarding patient status. It can be difficult for responders to quickly determine evacuation status due to the sometimes staggered physical placement of patients across an MCIs geographic area, particularly if those patients are not mobile and need ER support to move. Furthermore, as triage happens in waves with patients exchanging hands rapidly and often, consistent communication is difficult regarding concerns like administered care and last checked status. Each successive wave of triage builds upon previous iterations. Later waves also allow responders to spend more time with patients, gather more detailed information, and provide more definitive care. More specifically, first-wave responders spend around less than two minutes with each patient to ascertain rough estimates of patient status. ERs then attend to patients in triage order to provide care beyond basic life-saving interventions and gather further information. Thus, continuity of care is the primary concern in relation to consistent communication across triage waves.

Though current paper triage tags are immensely helpful, we deduced that AR could amplify their effectiveness by recording, communicating, and updating patient status (automatically and via responder interactions), supporting stunned responders limited ability to recall triage process details, decision support via data aggregation across multiple patients, demarcating and avoiding hazards, providing

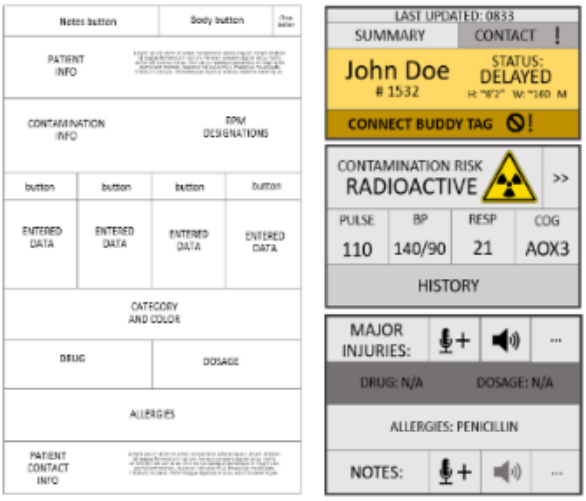

Figure 2: A comparison of an earlier iteration, lower fidelity wireframe (left) and a later, higher fidelity wireframe (right).

evacuation progress summaries, furnishing resource allocation status, and for training.

There are nuanced interactions between responder types and AR needs. For example, incident commanders need a more holistic understanding of triage efforts than novice and expert responders on the front lines. Our initial version of ARTTS centers on aiding front-line responders, particularly novices. However, expert responders also stand to benefit from ARTTS through its support of consistent communication and enhanced capacity to track evacuation and triage status. Finally, as SMEs identified novice responders as susceptible to becoming stunned, novice responders require more molecular, specific AR support.

SMEs indicated that any AR system will need to be at least bare-handed (with hands-free options preferred), usable in day or night conditions, easy to use, reliable, and create a shared digital environment for all responders. *Bare-handed* means that responders do not need to physically hold a tool to use the interface.

A *shared digital environment* means that responder A should be able to see and update a triage tag created by responder B, and that an incident commander can view a summary of triage information generated by responder A and B (and more). Additionally, AR-based triage must afford enhanced informational continuity across triage waves.

A *triage category* refers to both an assessment of the patient's current medical condition and delineates the order in which patients must be attended to and evacuated. Best practice standards utilize the Sort, Assess, Lifesaving Interventions, and Treatment/Transport (SALT) framework [37], so ARTTS does as well. Black refers to deceased patients, grey is for patients identified as expected to expire, red for patients who are in critical condition that can potentially be saved but need medical intervention in a hospital, yellow for patients who can receive stabilizing care in the field but will need hospital follow up, and green for patients experiencing minimal injuries (if any).

Visual elements leverage the Mixed-Reality Took Kit's (MRTK) standard dark blue backgrounds and white text seeking to mitigate limitations of the AR display's additive light model [56] and afford visibility in brighter conditions, as well as aesthetic consistency across tools. We used Nielsen's heuristics [57] to anchor our interface design efforts to ensure that ARTTS has an aesthetic and minimalist design, flexibility of use, employs recognition over recall, prevents errors, and exhibits internal consistency [57]. We also used gestalt principles to chunk information groupings within the virtual tag to ease visual search [58]. Moreover, we greatly expanded the salience of the assigned triage category by highlighting the entire informational box containing patient name, thus creating an emergent feature and supporting parallel search [58]. Flexibility of use was further attained via multiple interaction types, basing the virtual triage tag on the physical tag, and by providing access to process flow charts.

Our design efforts resulted in three main (but four total possible) interaction techniques: poke, ray-cast, and voice, with touch-pinch used for special interaction (e.g., grabbing the side of an interface window to 'physically move' the window location). These give responders flexibility of use redundancies and afford bare-handed use. Interactions can be used to select items and toggle between the triage tag and the SALT walkthrough interfaces. As ray-casting and poke interactions will still require responders' hands, voice provides a truly hands-free interaction option where possible. The range of interactions also ensures ARTTS elements can be utilized in near menu—within arm's reach—and distant menu contexts when necessary. Gaze was considered for this purpose as well, but early pilot testing revealed inconsistent effectiveness, thus violating the requirement that the system be reliable.

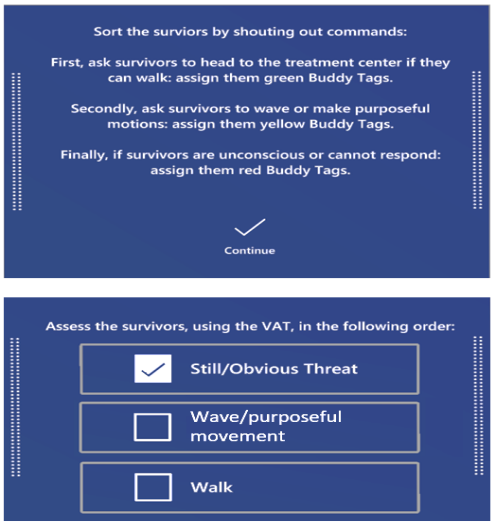

Figure 3: Initial sorting prompts. The first prompt (top image) is used upon arrival to the scene, while the bottom image acts as a prompt and progress checklist.

All ARTTS elements can all be easily moved and resized, which is the only way the touch-pinch interaction is currently utilized. ARTTS elements can 'follow' responders or be pinned to the physical world. Follow refers to a body-relative AR user inter-face element, meaning that the tag is anchored to the responder's body. Pinning the tag toggles ARTTS into a world-relative AR user interface element, anchoring ARTTS into a physical world location [59]. To summon any ARTTS tool responders need only hold up their palm with fingers closed in a flat position (leveraging MTRK's hand menu features).

In the following sections we present ARTTS' three main AR tools in the order we would expect them to be used in the field. All tools can be accessed via the 'toolbox' featured in figure 4.

2.4.2 Initial Sorting Prompt

SMEs indicated that arrival to an MCI scene can be overwhelming. Moreover, the sheer scale of the chaotic scenes can stun responders into inaction, as stressful events can potentially alter memory retrieval [60] and result in cognitive impairment [61]. From our discussions with our SME and the literature in training, we know that repeated study of materials mitigates retrieval inhibition the least, with repeated retrieval practice resulting in more enduring memories and enhanced retrieval [60]. Meaning, novice responders are potentially unable to retrieve critical process details quickly and accurately as the only real mediator for this effect is previous experience.

Stakes are high in MCI response with each second being critical. Thus, to improve response efficiency and support novice responders we created an on-demand AR-based initial sorting prompt (figure 3).

This prompt is uncomplicated and only utilizes the very first step of the SALT triage process. The instructional language is simple, and a check list is used to track progress. Initial sorting can be an initial kick to reorient responders back to the triage process. The interface is deliberately simple with outlines around each selection option to aid visual scanning [58].

The initial sorting tool employs poke, ray-casting, and pinch interactions. To ensure consistency, it uses the same aesthetic as all ARTTS elements.

2.4.3 Virtual Assessment Tool and ARTTS Tool Selection and Management

Sometimes a simple early kick is not enough. Should an ER need more aid they can leverage additional support via the Virtual Assessment Tool, or VAT.

VAT serves as an interactive alternative to the current static SALT flow chart and provides an AR-based "click through" experience to yield a decision tree that would otherwise need to be internalized (and recalled) by responders under duress. By design, the VAT complements the Initial Sorting Tool, by supporting the "ALT" processes defined by SALT (with the Sorting Prompt Tool supporting the "S" in SALT).

As responders select items in response to VAT's interface prompts (see figure 4, top image, where top two prompts are selected), the system tracks the appropriate paths through the SALT flowchart, eventually resulting in a final triage assessment. The assessment outcome then automatically populates a new virtual triage tag for the relevant patient with the appropriate data (leaving other areas like allergies blank for the time being). We designed this experience to minimize the number of interactions necessary. Though we could eliminate all interactions through a show/hide interaction by simply presenting the entire SALT flowchart as a static image, we felt this would increase mental workload and visual search time resulting in non-adoption of the tool. Additionally, we felt this was another opportunity to leverage gestalt principles as well as AR's visualization and interaction capabilities to improve existing processes and artifacts.

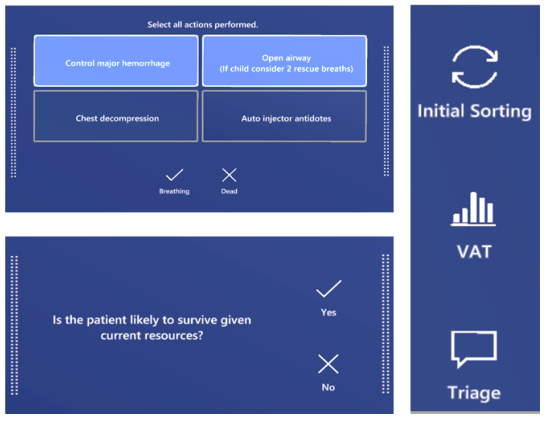

Figure 4: VAT prompts and tool manager. Top and bottom images are sample prompts provided by VAT. The right image is the overarching tool manager.

2.4.4 Virtual Triage Tag

We designed ARTTS' virtual triage tag [62] using a physical tag work artifact (figure 5) created by Evacu-Aid [63] thus ensuring user familiarity by migrating design elements to minimize upending of current processes. Further, the virtual triage tag adheres to current best standards of practice (SALT). The virtual tag (figure 6) is chunked to minimize visual search time [58] into three major sections found in the left column: general information, physical condition, and additional information.

General information provides characteristics of the patient such as name, height, weight, as well as overall triage status and time of last assessment. Overall status follows traditional naming and color coding to allow ERs to quickly identify patient condition (once specified by previous waves).

Physical condition data is gathered on responders' first inspection of patients. The triage tag supports collection of pulse, blood pressure (BP), respiration, cognitive status, as well as injury type and location if applicable as these areas are core to triage category delineation. We used insights from our expert interviews to define the specific quantitative values employed for status delineation.

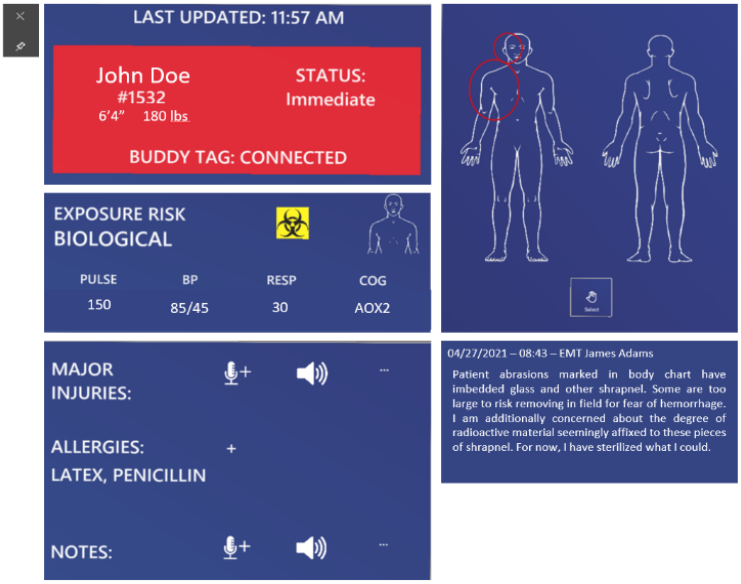

Figure 5: ARTTS' fully expanded AR-based virtual triage tag

Lastly, *additional information*, allows responders to elucidate qualitative details on patient injury severity that cannot be easily quantitatively captured such as critical information needed for effective medical intervention including major injuries, allergies, and other notes. Allergies creates a quick check for critical allergies identified by SMEs such as latex, penicillin, and sulfa. The major injuries and notes sections host voice recordings created by first wave responders completing early assessments. These voice recordings are available for aural playback and can also be dis-played as text through voice-to-text. Each entry in the tag is timestamped, while voice-to-text entries are additionally stamped with responder identity (responders log in to each headset, thus allowing for their entries to be coded with their information).

In the virtual triage tag, triage category is presented in the first chunk (instead of at the bottom like the physical tag), and the triage category color coding fully surrounds critical demographic information for patients. Responders can quickly delineate the assigned triage category by the color saliency, but status has also been redundantly encoded with explicit status labeling in the same visual area. As black appears transparent in optical see-through AR, we used two distinct shades of grey to represent the printed tags black and gray colors. Any potential problems perceiving differences between the two shades of gray is minimized by the aforementioned capacity to complete parallel search due to dual encoding [58].

Figure 6 features ARTTS' fully expanded virtual triage tag. The items in the right column, the body chart and voice-to-text boxes, can be separately toggled open or closed by selecting the small person-silhouette and major injury ellipses respectively shown in the main tag (or by saying "show/hide body chart"

and “show/hide major injuries/notes”). When opened, the right-side panels slide open to the side of the main tag instead of replacing the main tag. This was an indicated preference found through design walk-throughs with our primary SME that affords all data is visible when desired, and collapsible when not.

Instead of using a pen or pencil to mark physical areas of concern and injuries, responders can use ray-cast or poke interaction techniques to mark up the virtual triage tag’s body chart. When marking the body chart, responders can toggle between a circle and eraser tool. The circle tool allows ERs to place a circle (allowing for customized placement) onto any part of the body chart. Early designs included an eraser tool to remove any circles with one poke action, however the modal interface was found to be more cumbersome than simply allowing users to simply re-select existing circles to remove them.

Figure 6: Participant using the virtual triage tag.

Pulse, blood pressure, respiration, and cognitive data can be entered with voice or via poke and ray-cast interaction with a virtual keyboard. Future iterations will limit the numeric data areas to a numeric key pad. Finally, the digital nature of the triage tag allows for bidirectional patient status changes. The traditional paper tag only affords category changes if the patient worsens, while the virtual tag allows for patient status improvements to be documented as well without losing the patient’s category assignment history.

2.5 Triage Tag Evaluation Results & Discussion

Eleven emergency responders (3 female, 8 male; Age range 20 – 67, avg. 43.8) participated in a usability evaluation of the triage tag, the most complicated element of ARTTS, so that we could ascertain the virtual tag’s usability and associated mental work-load. None of the responders had any experience in AR and ten had experience with physical triage tags. UX studies have been found to not need large sample sizes, with Rex & Pyla describing that empirical evaluations with at least three to five users can often identify 80% of usability issues within an system [46]. Equipped with this, we feel that eleven participants are sufficient for our initial purposes.

Participants were first given AR user interaction familiarization tasks designed to train participants on poking, ray-casting, scrolling, and moving the tag. Participants were trained in poke interactions by being asked to poke a random layout of ten virtual boxes labeled 1-10 in ascending order. This was done 8 times, yielding 80 pokes. This process was repeated for ray-casting. For scrolling, participants were asked

to scroll through a list (in near range, so like a 'poke touch' interaction) to find a specific fruit and then click on it when they saw it, yielding 8 scroll tasks. Finally, participants were asked to ray-cast a cube from one location to another within a designated cube (far field), yielding 8 object movement tasks.

Upon completion of familiarization, participants are then given either a paper or AR triage tag (order of which was counterbalanced across participants) and walked through a set of benchmark tasks built such that all interactions are completed more than once and to give participants a good overview of tag usage. Example tasks included "add latex to the allergy list", "record a pulse of 120", "update the patient status", etc. These benchmark tasks varied only in the specifically requested interaction styles to achieve the task.

After completing the benchmark tasks, participants were asked to fill out a System Usability Scale (SUS) survey [64] and a Raw Task Load Index (RTLX [65]—NASA TLX [66] without pairwise comparisons to mitigate survey fatigue) [65]. Then participants were given a new set of benchmark tasks, comparable in difficulty, on the remaining tag modality, with the same surveys completed afterwards. The AR triage tag did have some unique questions associated with it. The first was the short form version of the User Engagement Scale [67] used for application assessment. The second was a self-reported assessment of their comfort in their AR interaction skills (1-10, 10 being the most comfortable.), as well as whether they felt their AR interaction skills improved with time (yes or no).

IV	**DV**
Virtual vs. Paper	SUS
	TLX
	--
	UES (AR only)
	Comfort Report (AR only)

An initial assessment of data normality showed some concerns for overall SUS scores. A continuous fit evaluation displayed a right skewed (-2.21) curve with a leptokurtic kurtosis (6.077). These scores are both well outside what is considered a normal range of +1 for each [68], indicating non-normality. Interquartile range (IQR) evaluation showed one score well below the Q1 cutoff of 65. Huber M-Estimation was also used to estimate center and spread of this data with a default K Sigma of 4, ensuring an appropriately conservative evaluation of the data. Huber's evaluation also pinpointed the same datapoint as an outlier. Thus, this outlier was excluded from our analysis. This exclusion yielded a new skewness (.13) and kurtosis (-.48) which is centered within that aforementioned +1 normality range.

Overall, ARTTS' virtual triage tag received an average SUS score of 74.5. This is above the recommended average of 68 [69], thus simultaneously assuaging any immediate major concerns with the

interface while also revealing there is room for improvement. The paper tag had a higher SUS score, but a matched pairs evaluation found this difference to be statistically insignificant.

Regarding mental workload as measured by the RTLX (out of 21): mental demand (9.8), physical demand (4.7), temporal demand (8.2), and frustration (5.5) were all below 11 (or moderate) for the virtual tag. Physical demand was the least concerning with a low average score of 4.7 (out of 21). Conversely, the biggest concern is that participants reported an average score of 15.5 (out of 21) for their perceived failure in accomplishing the task using the virtual tag. The paper tag had non-significantly lower results for the overall TLX score and across all dimensions except physical demand ($p > .01$), which is not surprising given that our tested tag required hands interaction.

Finally, we administered a short form version of the User Engagement Scale [67] to provide a deeper evaluation of the virtual tag and overall experience with AR. With 1 being strongly disagree and 5 being strongly agree, the subscale of perceived usability has a mildly positively leaning average of 3.67, further supporting the SUS result that the virtual tag has average usability while still having room for improvement. Focused attention had a similarly middling score of 3.33. More positively, aesthetic appeal of the tag yielded an average score of 4, with a perceived reward factor of interacting with the tag receiving a 4.67. All participants reported feeling highly confident in their AR skills and like their ability to interact in AR got better over time.

In summation, the virtual triage tag has held its own as compared to a physical, paper version of the triage tag in terms of user perception in the context of our study. I.e., the virtual tag did not significantly negatively diverge from the paper tag as indicated by our SUS comparisons. Though any change inherently yields a degree of process upending, these non-significant differences in SUS scores speaks to a degree of success in our goal to minimize upending existing triage processes. However, future iteration is warranted to maximize the tag's effectiveness and to minimize any negative tradeoffs with the paper tag. The core areas for future focus will be on the physical fatigue facet and in helping provide user feedback to indicate success in tag interaction.

2.5.1 Design Principles

From a qualitative perspective, we noted several design principles for AR in MCIs, as well as UX findings that may be of interest to other AR practitioners.

Responders are sifting through several information streams and have to move quickly at first, eventually slowing in subsequent waves. Further, ERs are likely in dangerous environments when actively responding to an MCI. This means that **AR applications should minimize visual search, afford different response strategies** [70]**, and not subtract from ER situation awareness.** These can be achieved using

gestalt principles like common region, ensuring applications allow for a range of fidelity in each use, and by minimizing the graphical footprint while providing quick toggle on/off of said graphics. Our triage tag uses *common region* to chunk similar information and input types. Further, the tag has some quick response data (simple numbers) and longer response data (verbal notes) to support shifting ER strategies.

Graphics and applications should leverage Nielsen's heuristics [57] to further facilitate efficient and pleasant use, while critically ensuring reliability of the system through error prevention and easing recovery from errors. Designers should further **consider the selected hardware's pros and cons carefully.** We chose a HWD based on ER feedback and what we felt were worthy trade-offs for a feasible future of AR. **Hardware should be reconsidered on a case-by-case basis** for the intended use case. Finally, **graphical elements should be robust** enough to handle as wide a range of bright environments as possible, with visual redundancies for any colors that may be washed out on either end of the light spectrum.

More specifically, participants found poking much easier than ray casting, suggesting that **user interfaces should be placed closer to users for easy access.** Our triage tag is representative of likely the largest single UI that can be effectively used quickly via touch without more detailed practice and training and given today's AR head-worn displays' field of view.

Second, we noticed that scrolling through menus and large fields of vertical text (e.g. notes) proved to be difficult for some participants. We observed that some participants engaged in an "exaggerated scroll motion", that is, physically lifting their arm above the head in to accomplish an effective vertical scrolling. ERs need reliable and fast interactions. Further, we observed that during scrolling or selection at distance, both of which used a pinch, participants sometimes did not separate their fingers enough and would lower their hand before the "pinch release" was recognized by the system (resulting in undesired continued scrolling or an untended non-release of a UI selection, respectively). Thus, **practitioners may consider a "paging" mechanisms instead (e.g., buttons with "next" and "prev") to poke through subsets of lists or text entries instead of scrolling a continuous list.**

Participants also found numerical data entry on the standard HoloLens/MRTK keyboard tedious, which suggests that a numpad (and redundant voice capability) may be a more effective input mechanism when there is a need to specify numerals. In a similar fashion, we also noticed that for entering blood pressure, we needed an **input mechanism that matched the domain-specific ecological mental model** [71] for blood pressure readings. So, for example, instead of entering systolic blood pressure and diastolic blood pressure as separate fields, we added a "/" to the numpad to allow for blood pressure to be entered as "120 / 80". Participants also did not re-position the virtual tag as much as we thought they might (i.e., in order to see the virtual tag better or to poke interact) and instead moved their body to more convenient viewing locations. Lastly, the HoloLens would often lose tracking when participants placed their head down

low to the ground and turned their head to the side in order to listen to a supine patient breathing. Upon returning their head to a normal position, tracking would resume however the ARTTS' UIs would sometimes shift to a slightly different position. **Practitioners should be wary and careful about ensuring tracking is consistent and reliable.**

2.6 Conclusion

Our review of literature, training materials, and expert interviews identified several opportunities for AR support in emergency response. SME priority rankings for MCI triage guided our team to focus on recording, communicating, and updating patient status, supporting stunned ERs, and decision support.

Our analysis of interview transcripts suggests that AR systems for ERs need to support bare-handed interaction when possible, minimize the need to use hands in general, be visible in day and night conditions, are easy to use, are reliable, and create a shared digital environment to support multiple user roles (ERs, incident commanders, and in the future healthcare professionals receiving patients at hospital facilities). The team identified optical see-through AR as ideal to ensure reality is truly augmented (not replaced) and situation awareness is preserved.

Our current triage tag received middling to good scores on most tested facets. Though the paper tag trended more positively across these dimensions, these differences between the tags were not significant except for in relation to physical demand.

ARTTS supports standard recording of patient status similarly to the original physical tag. However, ARTTS data recording is augmented by the ability to update the tag and track provenance of updates, and the ability of the virtual triage tag to be accessed by any ER at any time and location. Moreover, the option to send data to hospitals in advance is an additional promising improvement that we hope to explore in future work. ARTTS global sorting and VAT interfaces further augment the ER experience by providing prompts to guide early triage requirements.

2.6.1 Limitations

Regarding interactions, ARTT'S uses speech-to-text which has its own limitations, especially in the context of loud environments like what might be found in MCI's. The HoloLens 2 allows for shifting of the sound sensitivities to afford better speech command intelligibility, but these may still be overwhelmed depending on the MCI. Additionally, standard problems with AR in bright, sun-lit outdoor settings are still present. ARTTS does not currently provide a fully hands-free alternative to move the interfaces panels in space (one must touch-pinch), thus requiring the system be in the near field and only moved within the near field.

Further, our current study focused on physical interactions and did not include voice commands, which may have been a mitigation strategy for fatigue.

We concede that there are limitations to HWDs in relation to their feasibility due to limited battery life, network connection, and limited effectiveness in outdoor and bright settings. However, mobile devices have long been available to responders and are outlined as not feasible via SME interviews. As HMD technology continues to develop, we foresee them quickly surpassing mobile devices in contexts where users need their hands free.

This work includes a small sample size of only eleven. While we feel eleven was sufficient for our purposes, more is preferrable. Further, participants were limited to one session of using the triage tag. As all participants indicated that interactions got easier with time, some perceived usability issues may be a result of not having enough familiarity with the tag itself because while it is heavily inspired by the original paper tags, there are still some changes that would require re-training if used in the field. Further, while our virtual tag did not negatively diverge from the paper tag, we concede this also means that these comparisons do not make clear any obvious 'gains' to be had from implementing AR in the field as the scores for the virtual tag were not significantly more positive. Simply replacing the paper tag with an AR version does not inherently yield meaningful change. We believe, however, that the gains of our tools are more likely to be realized else-where. Namely, in the field vis-à-vis additional and dynamic systems support offered by our virtual tag that are not available via paper medium. Our virtual tag could yield improved evacuation efficiency, more effective patient triaging, better resource management, and many other higher order benefits not evaluated by this study.

2.6.2 Future Directions

Moving forward, we envision ARTTS being integrated with ER-based system architectures such as Panacea's Cloud [3] or a Cloud-Fog [17] to support possible functions like the provision of advanced hospital notice. More work is needed in terms of in-situ usability evaluation strengthened via formative user evaluations in realistic MCI triage settings to obtain quantitative measures of effectiveness relative to the traditional paper triage tag and mental recall required for initial sorting and SALT. Studies regarding technology acceptance, VAT access versus strictly recall during response, and the effectiveness of initial sorting in reorienting stunned ERs should also be completed.

Future work in the AR triage space should not only build upon the areas focused on by ARTTS, but also consider possible solutions for the other identified use case areas not touched on by ARTTS such as wayfinding, hazard avoidance, evacuation progress summaries, resource allocation areas, and training.

Further, future work should endeavor to include more participants to ensure normality can be better assessed. Any future studies should include more training on the triage tag itself (not just on AR interactions) to ensure there is not a confounding effect with learning the new orientation of tag elements. The AR triage tag appears to need further iteration to bring the usability from 'average' to 'excellent' and to improve users' perceived performance. Finally, it is critical to assess whether these non-statistically significant differences have large practical implications, particularly in the context of actual field work. Future evaluation should consider ARTTS in at least a more intensive training scenario. This is particularly true as some of the core benefits of ARTTS that might make it favorable to a paper tag despite increases in some facets of mental workload were not well displayed in a controlled evaluation setting (like how paper tags are damaged by rain and mud, etc.). In short, as with many emerging applied AR endeavors, the ARTTS UX and associated hardware need to be stress tested in more field-like scenarios to better prepare it for adoption by ERs. Finally, we intend to evaluate the higher order gains of ARTTS like improved evacuation times and more efficient triage processes.

3 Study 2 – A Cognitive Work Analysis on Mass Casualty Incident Triage: Lessons Learned and Design Considerations for Future AR Support Systems

3.1 Abstract

Mass casualty incidents are complex sociotechnical systems comprised of several complicating characteristics, the most difficult of which to deal with design-wise being the high degree of disturbances within the live-response domain. This work offers a cognitive work analysis on a subsystem of mass casualty incidents: Triage. This is because triage is highly time-sensitive and imposes massive informational and cognitive requirements on responders. We also hope insights gleaned at this lower system level could propagate and inform insights at the larger, broader MCI level. Further, this work highlights the need for systems to support responder situation awareness and the importance of actor communication in attaining that situation awareness. There is a burgeoning interest in infusing augmented reality (and other computer-based technologies) into the emergency response space. Thus, we feel it prudent to evaluate this sensitive and volatile domain of MCI to proffer design implications for how prospective AR assistive systems can meaningfully be injected.

3.2 Introduction

The World Health Organization defines mass casualty incidents (MCIs) as events generating more patients than available resources can manage with routine procedures [1] and require coordinated, rapid emergency response [6]. The public expects MCIs and other incidents to be effectively managed by emergency responders independent of the context in which the MCIs occur [3]. Unfortunately, case studies have revealed that many crises are poorly managed [3], [4] with shortcomings associated with leadership, planning, communication, and resource management [2], [3].

The WHO further indicates that MCIs require "extraordinary assistance" [1], with there being "significant gaps" in the science surrounding MCIs [37]. To assist the burgeoning endeavor to support MCI responders [72], [49] through the development of this extraordinary assistance via augmented reality (AR), this work seeks to provide a Cognitive Work Analysis [73] as outlined by Vicente's seminal book [33] and Kilgore's sample application in the context of air traffic control [74]. This is because "… a thorough work analysis of some type must be conducted if computer-based systems are to effectively support human work in complex sociotechnical systems" [75], with each phase being nested in the prior phases to abstract constraints and delimit the true degrees of freedom available for action [76]. Given the breadth of complexity of MCIs, this work will take a similar approach to Kilgore [74] by delineating a key component of MCI response to focus on: Triage.

Triage is a subsystem or subtask of MCIs subject to many of the same characteristics as the overarching MCI domain. Triage is the assessment, treatment, and assignment of evacuation priority of injured persons [34]. Poor triage coordination leads to increased mortality [77] through the resultant delays in response [36]. This delay in triage is detrimental as it can exceed the 'golden hour' of trauma. This golden hour is a crucial time when treatment greatly increases survival rates [7], [5].

AR is a burgeoning technology in emergency response, triage, and MCIs [5], [7], [35], [36], [38], [39], [42]–[44], [78] with numerous use cases [25] due in part to AR's ability to superimpose information atop the real world [26]. However, given AR's novelty paired with the volatility of MCIs, penetrating the adoption barrier is difficult, with technology acceptance noted as a "consistent hurdle" within emergency response [25]. At the time of this writing, there are no AR-based wide-scale adopted tools for MCIs or triage. None of the aforementioned works that leverage AR ([21]–[30) focus on the nuanced undercurrents of the *system* of MCIs or subsystem of MCI triage, instead focusing on their explicit surface application area, typically via an approximation of a usability evaluation. We posit that designs considering the findings of a full Cognitive Work Analysis (CWA) will have better alignment to the demands of the work domain, introduce less disruption, and will be closer to "get[ing] it right the first time around" [33]. Thus, we pair this CWA with AR-specific design considerations informed by both the CWA analysis and by lessons learned in our own development of an AR Triage Tool Suite [25].

3.2.1 MCIs as a Complex Sociotechnical System

Emergency response to MCIs constitutes a pop-up complex sociotechnical system. We define this space as 'pop-up' as the events themselves are disturbances to standard society that are highly comprised of other disturbances (factors not anticipated by designers that can impact the system [33]). More specifically, MCIs feature ten of the eleven characteristics of complex sociotechnical systems [33], [79] namely MCIs are/suffer from:

- Large Problem Spaces
- Social
- Heterogeneous Perspectives
- Distributed
- Dynamic
- Hazardous
- Coupled
- Uncertain
- Mediated Interaction
- Disturbances

A complex sociotechnical system refers to social and technical subsystems constantly interacting with the environment [79]. The environment can impact the system, and the system can impact the environment. These complex systems do not need to rate highly on all eleven possible characteristics but will usually comprise several dimensions and are marked by the need for workers (responders) to deal with the complexity in real-time [33].

MCIs are varying but often large problem spaces [80] as they are composed of many potentially relevant factors, including thousands of diagnoses [33], medication on hand, available ambulances, the crisis's scale, the crisis's environmental spread, etc.

Effective MCI response is highly social as it requires many people to work together with clear communication and effective coordination [3] distributed across several locations or large areas. The difficulty associated with this coordination is compounded by the heterogeneous perspectives resulting from responders hailing from several discipline types (police, firefighters, emergency medical services, hazmat, SWAT, etc.) [37]. Effective MCI response further involves the coupling (interaction) of many subsystems associated with crisis mitigation, like evacuation, setting up medical tents, clearing hazards, search and rescue, etc.

There is profound uncertainty in the context of MCI response, particularly concerning these coupling interactions, due to the impoverished input [33] of the overall crisis because relevant data comes from all sides and is not easily aggregated to form a complete picture of the crisis. Due to the nature of MCI response often involving human patients and evolving hazards, there are mediated interactions in this space as there are many times when elements within the system cannot be observed unaided (blood pressure of patients [33], carbon monoxide leakages in the building, etc.). Concerning hazards, there is often a high degree of risk imparted upon emergency responders in their participation in this pop-up complex sociotechnical system. Responders could be subject to radiation exposures, chemical exposures, etc., as they attempt to navigate the MCI. Inappropriate actions can have catastrophic consequences for the responders, the overall response effort, patients, and property [3].

MCIs are dynamic [80] in that the event can take a variable amount of time to react to responder intervention (i.e., a shooter can be disarmed within minutes, and the crisis ends, or it can take weeks to put out a forest fire).

MCIs are unanticipated events or disturbances constructed by a confluence of disturbances. MCIs differ from standard emergency response because MCIs are low-incident events with massive consequences. Standard emergency response procedures denote a much smaller scale complex system that often involves only some of the abovementioned characteristics. In addition, MCI rarity means that many responders have limited experience necessary to develop skills to mediate some of the difficulties of MCI [81]. Further, though rural responders serve 75% of the United States, rural responders have fewer certifications and resources to attain these certifications [82]. The responder's lack-of-experience concern further results in questions about whether responders can learn from failures occurring in MCIs to ensure those failures do not reoccur [3].

Standard emergency response events are marked by fewer disturbances. This is because standard response events are more predictable and, therefore, more 'designable.' This yields unique difficulty in affording effective MCI designs, especially given that MCI risk means designers must "get it right the first time around" [33]. The stakes are so high for both responder safety and public safety that designers do not benefit from iterative (or evolutionary) design. MCIs are an exemplary case of designers needing to create a system that is "… capable of functioning effectively or especially under rare, unanticipated events" [33]. MCIs can consist of patrol and suspect neutralization via SWAT [49], search and rescue, evacuation [48], fire rescue, triage, hazard detection, and much more.

Given the complexity of MCI triage response and the environments in which they occur, and the promise of AR, the objective of this chapter is to *develop a deeper understanding of the MCI triage work domain and how that interfaces with AR design.* From here, we ask the following research questions:

7. What does a cognitive work analysis (CWA) reveal about AR's potential to improve for MCI triage?

8. What are the specific design considerations (if any) that can be gleaned from the CWA for AR in MCI triage?

This chapter will outline each of the five phases of a CWA. Each phase will have an explanation of what that phase captures within the subdomain of triage in MCIs and the design implications that come about from that phase of the analysis. Our goal is to provide design implications for ecological interface design that can be used in tandem to a user centered design process. This is because an ecological design process focuses on the work domain and a user centered design process focuses on the end user. We posit that both are needed to design for high-risk spaces like MCIs. This work is heavily informed by our previous work in this area and information gleaned from subject matter experts over the course of conducting that previous work [25].

3.2.2 Related Work

Stephan et al., provide a cognitive work analysis of augmented reality in marine navigation [83]. Most critically, they note AR's capacity to superimpose context-relevant data onto the work domain, thus facilitating observations and task definition (further discussed in the MCI context in phase 2 of this chapter (the control task analysis). Clark et al., note the importance of understanding cognitive systems for emergency room MCIs due to the intense time pressures and high acuity of patients exacerbating cognitive workload during critical decision making [84]. However, to our knowledge, there are no other works providing a cognitive work analysis on mass casualty incident triage in the field, let alone as a means to facilitate AR system design, though Feigh et al., do outline that adaptive systems (an emerging area of AR support)

require a complex understanding of the user's environment, task environment, and cognitive state in order to be effective—all of which are worked toward via the following CWA [85].

3.3 Phase I: Work Domain Analysis (WDA)

The first phase of the CWA is the Work Domain Analysis (WDA) [86]. This phase is used to identify what information needs to be measured and derived and how that information should be organized [76] in relation to the work domain's functional purpose and internal constraints [74]. The work domain (field of activity) is described by leveraging an abstraction hierarchy and mapping the findings in the abstraction decomposition space (ADS). A sample containing the most important information for our interpretation is provided in Figure 7 The remaining two levels are the Physical Function (e.g., gloves, emergency medication), and their underlying Physical Forms (e.g., latex or nitrile, blue, purple, green, black, or white; Epinephrine, Narcan, Albuterol, Ointments, Morphine). While important to understanding the system, in the context of MCI triage, these lower two functions serve more as requisite resources. Those considered in our analysis but not featured here are are pictured in the appendix (see Figure 31).

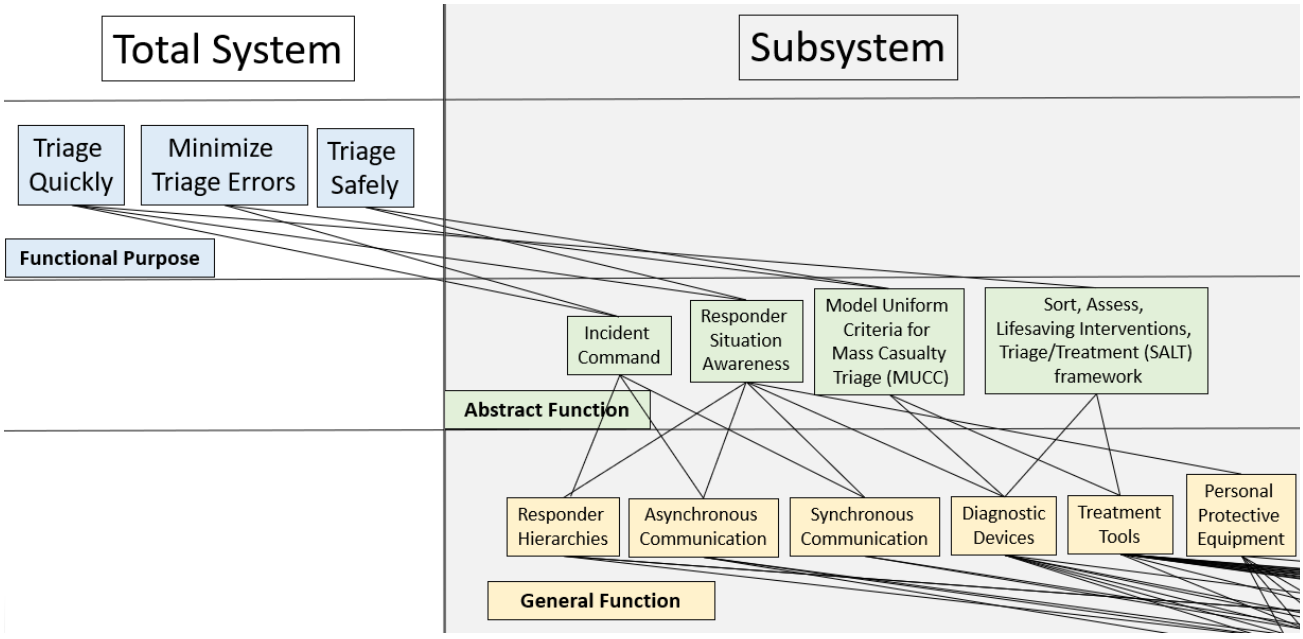

Figure 7: First 3 levels of the ADS for Triage

We identified three functional purposes of the work domain: to perform triage activities quickly, safely, and to minimize errors in that process. To that end, four abstract functions (via a priorities and value approach) were identified:

- Incident command plays a crucial role in ensuring triage is done quickly. Incident commanders act as the hub through which disparate information is connected, synthesized, and relayed.
- Responder situation awareness is crucial in understanding all relevant responder traffic, terrain, civilians in need of assistance, already aided civilians, hazards, weather, and available resources for allocation.
- The Model Uniform Criteria for Mass Casualty Triage (MUCC) framework outlines requisite components of effective MCI response that responders must be responsible for knowing and executing. This is a national standard for MCI response of skill requirements [51].
- The Sort, Assess, Lifesaving Interventions, Triage/Treatment (SALT) framework is similarly something responders are responsible for knowing and executing correctly. This is a national standard for MCI triage [37].

The identified physical functions not depicted in Figure 7 denote the exact tools leveraged by emergency responders, like gloves, oximeters, tourniquets, etc. The physical form (also not depicted in Figure 7) outlines properties of these tools, i.e., gloves are latex or nitrile, oximeters are ~3 inches long and made of plastic with an onboard CPU, etc. The physical function and form tiers quickly balloon into incredible scale; hence, they are not being pictured here. However, we have broadly classed these tools within the hierarchies outlined by the general functions: responder hierarchies, asynchronous and synchronous communications, diagnostic devices, treatment tools, and personal protective equipment. Generally, communication involves the usage of radios, flares, caution tape, and triage tags. Diagnostic devices include blood pressure cuffs, stethoscopes, etc. Treatment tools encapsulate sutures, emergency medications, etc. Finally, personal Protective Equipment denotes gloves, masks, eye protection, etc.

3.3.1 Design Implications – Phase I (WDA)

Future computer assistance must help triage occur quickly, safely (for both responder and patient), and minimize triage errors. If the designer's goal is to integrate their potential AR aid with existing triage practices, this theoretical system must not violate the domain's intentional constraints, meaning it should integrate the MUCC and SALT frameworks. These frameworks need integrated because they are federal standards built over years of best-practice research by subject-matter experts to maximize both patient and responder safety. Further, these frameworks are the standard strategies used by responders in the field and in their training, meaning that working to support these systems gives designers a generalizable foundation upon which to build. Responder situation awareness is a critical component of the MCI abstract functions

that further leverages the MUCC and SALT frameworks. Situation awareness is defined as the perception of environmental elements in a given time and space and the related comprehension of their meaning and projection of their status in the future [87], [88]. In the context of MCIs, responders must be aware of hazards, resource management, rescue efforts by other specialized responders, and various patient statuses as the MCI evolves. Functionally, future assistive devices need to work for several communication strategies and connect responders to each other and incident command. Incident command presently acts as the primary means disparate information is synthesized and disseminated and thus is integral to any future systems. Finally, due to the intense nature of MCI response, assistive technologies should at worse be 'neutral' in the event of a failure, e.g., if a device breaks, it should not leave responders off worse than they would be without using the device. For AR, this poses device-related constraints where perhaps video-passthrough AR is not feasible as a failure blacks out responder vision. Future supportive equipment must also consider personal protective equipment (PPE) and other field tools. E.g., AR devices cannot require responders to remove or cause misfit with PPE. Moreover, AR must work in tandem with diagnostic devices and treatment tools.

3.4 Phase II: Control Task Analysis (CTA)

Though MCIs are highly variable, there is still consistency in some of the procedures underlying triage expectations. Thus, the purpose of this phase is to better understand the requirements informing these recurring actions [89] to "… realize the domain's functional purposes" [74]. We are trying to derive what goals must be pursued, the associated constraints, and the relevant information [76] independent of the actor. To understand what needs to be done (control tasks), we have completed a decision ladder seen in Figure 8.

The ultimate goal of our analysis is to ascertain a responder's cognitive journey through a complete triage process and how their cognition interplays with the tools and environment of the scene. For our purposes, we have outlined the ultimate goal, which is to prioritize a balance between accuracy and speed, with one of our functional purpose goals. From there, responders must define the *task* based on the context. We have provided a sample case of 15 patients in all categories of the SALT framework. Again, the task is to complete this triage process accurately and quickly. To that end, the responder must conceive a response plan (i.e., *formulate a procedure*) based on the available resources, number of injured parties and the scale of their injuries, the physical environment, and other responder input. Then, they must *execute* this plan while having the requisite vigilance and understanding of relevant signals, signs, and symbols (*activation* and *alert*) that constitute a disturbance large enough to warrant a shift in their formulated procedure execution.

As responders complete their triage process, they must constantly *observe* the work domain for relevant informing variables to monitor system state or the current triage status. For example, responders must *identify* how many tourniquets are left, how many civilians have not been treated, etc. Of course, this must all be *interpreted* in relation to the *ultimate goal* of completing a triage process of all injured parties until the scene is cleared (no patients remain on the scene).

Experienced responders may utilize a shunt (denoted via dotted lines), jumping straight from the *interpretation* box (information processing routine) to a procedure generation circle (state of knowledge). This is because they do not need to consistently redefine their goal and task, whereas novice responders (often assigned grunt work) may need to shift their goals and tasks. More specifically, expert responders can jump straight from "interpreting the scene" to the requisite procedure without redefining the goals, tasks, and procedures themselves. Alternatively, novice responders may be afforded the shunt from "interpret" to "procedure" via direct orders from a more senior responder defining their tasks for them. Furthermore, expert responders may also use a shunt from the identified knowledge gained to the procedure step due to their expertise. The variance between these two shunts also captures the difference in response complexity. I.e., more complex environments will require experts to go all the way up to a deliberate interpretation stage, whereas less complex conformal relationships may become 'second-nature', and thus responders jump straight to the relevant procedures. These shunts indicate a synthesized understanding of the current domain status and an ability to predict the development of said status.

3.4.1 Design Implications – Phase II (CTA)

Though MCIs are highly variable and will have more activation-level disturbances than standard response scenarios, many opportunities exist to improve existing flow processes in interpreting the scene and then executing the requisite response. Particularly as responders need to be constantly vigilant, future designs should ensure they do not hinder situation awareness at a minimum, while ideal systems will facilitate greater situation awareness, particularly in helping gather ecological domain data and providing alerts in the event of sudden patient status shifts or hazards as informed by other responders on the ground. These systems should explore offloading some of the cognitive workload associated with the *interpret* step with conformal displays. AR is situated to facilitate this by augmenting the responder's reality. As with our original paper, the AR app provides 'cards' of patient statuses above each patient in order to offload cognitive interpretation and working memory demands of the scene [25]. Due to the rarity and severity of MCIs, future systems should foster the ability of novice responders to learn the expert shunts faster (ideally without sacrificing the requisite context-conditioned variability experience needed to be well-rounded responders) [90].

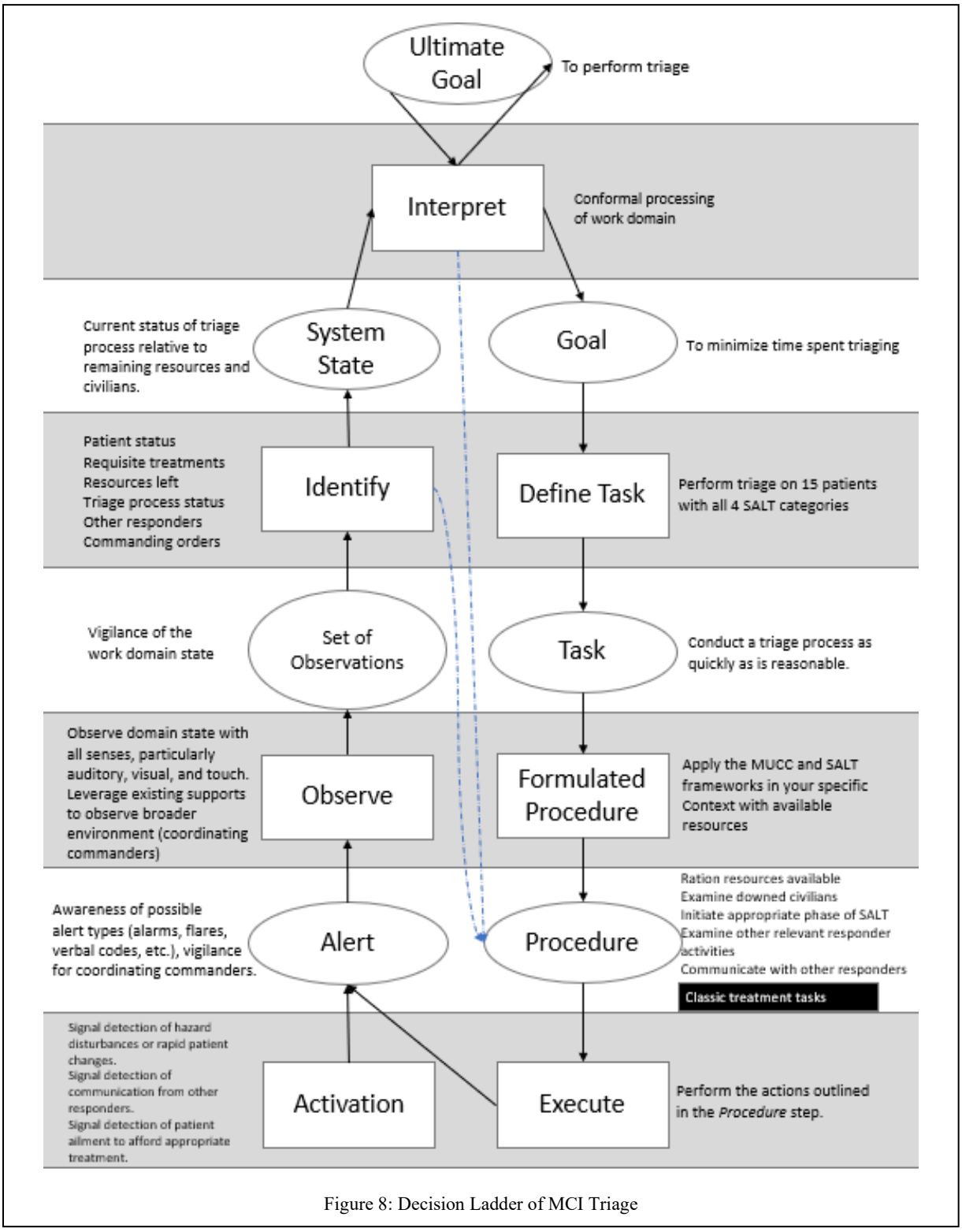

Figure 8: Decision Ladder of MCI Triage

Our original triage tool suite provides this in three ways: 1) responders can get a 'jump start' prompt to help kick start the triage cycle for stunned responders overwhelmed by the interpret box and how to define the next task. 2) Our tool provided a virtual assessment tool to work novice responders through the triage

process in situ, and 3) our virtual triage tag / 'card' view facilitates the support of other responders more directly. Thus, well-developed conformal displays helping bridge the synthesis of related but disparate data could provide novice responders with earlier access to these 'experience-based' shunts available to expert responders. Future systems should also ensure those facilitated shunts are meaningful for the experts themselves and perhaps provide higher-order synthesis, creating new shunts for experts.

Though the outlined control task does not analyze any interaction with any specific technology or support structure (like how to use the radio) or AR, AR is nuanced and variable. Thus, it did not seem a meaningful exercise to outline a prospective AR tool that has not been realized in the field given its lack of practical data and the wide array of AR's possible manifestations. Thus, instead, the control task outlines the tasks that provide the feedforward steps necessary to move a disaster toward a 'cleared' status. Responders require information (input) and then must engage in required actions (outputs) [89].

Our description of the MCI work domain did not include the eighth complex sociotechnical system characteristic: Automation. This is because there tend not to be, at present, highly automated systems that require human monitoring in emergency response, triage, or MCIs. However, in the proposing of conformal systems, it is possible and probable that AI and other automated systems will find footholds in this domain. Given that technological acceptance is an issue in emergency response, it is critical that future automated systems not add significant difficulties (unless they provide significant gains in other characteristics without compromising safety).

3.5 Phase III: Strategies Analysis (STrA)

The next phase of CWA is the strategies analysis [70] and intends to investigate different ways to achieve the same overarching control tasks (the 'how) [74]. Our strategies analysis strongly echoes the work of Sperandio [70][91] in that we discuss three strategies based on the number of patients (in Sperandio's case, planes) needing triaged and their relationship to the requisite cognitive load. See Figure 9: Strategies for MCI Triage for our strategies. Each one requires an understanding of the field status and results in an updated field status or a cleared field. Thus, all strategies determine rough triage categories for all patients and require general field information.

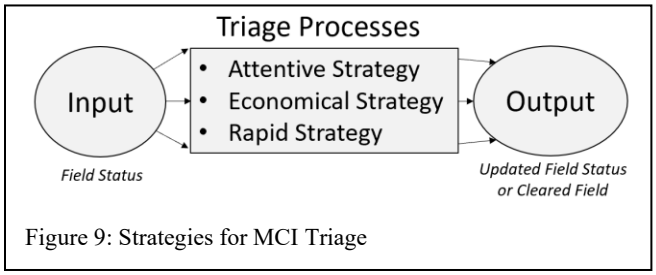

Figure 9: Strategies for MCI Triage

MCI triage leveraging the SALT framework happens in waves and categorizes patients into five possible triage categories: Minimal (green), delayed (yellow), immediate (red), grey (expected), and black (deceased) [37]. Based on a patient's category, patients will be seen immediately (red) or last (green). Green

patients have minimal injuries, yellow patients are more injured (will perhaps need hospital care) but do not need stabilizing, red patients need immediate stabilization, grey patients are unlikely to survive even with stabilizing measures, and black patients have already passed away. *Waves* mean that a responder will quickly group people based on severity and seek those who are severe but savable first (red). Once all are stabilized, a responder can double back to provide more detailed care. Alternatively, another responder can return to the patient if all severe patients are being tended to by others. The following strategies outline most specifically the "identify," and "observe" boxes in the aforementioned decision ladder, informing the volume of information dealt within the "interpret" box.

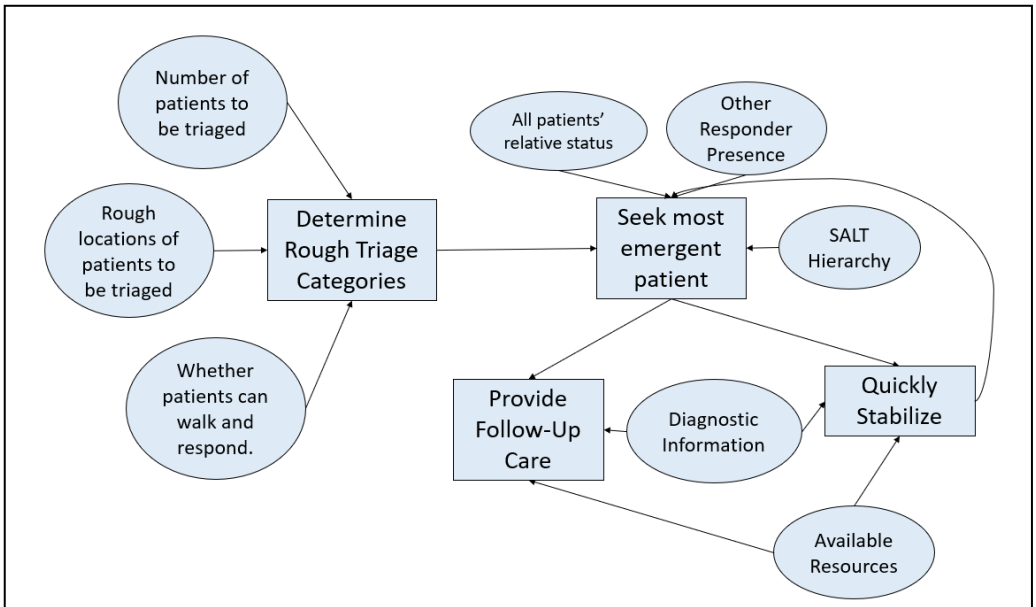

Figure 10: Information Flow Map for Rapid Strategy

The Rapid strategy (seen in Figure 10: Information Flow Map for Rapid Strategy) refers to when responders only work to stabilize an individual before quickly moving on to the next patient (like our decision ladder in Figure 8). Once everyone in need of stabilizing care has been evaluated, only then will responders circle back to those in less survivable critical condition (grey), back to red category patients for more detailed care and information gathering (name, age, etc.), or perhaps to yellow condition patients. Responders spend the minimum amount of time required with each patient to ensure they can stabilize as many patients as possible during their 'golden hour' of trauma. Category determinations are acquired much faster and have less granularity in their justification. I.e., simply being unable to walk immediately connotates a yellow to red assignment, even if ultimately the patient only has a sprained ankle. Furthermore, patients are treated automatically and to a minimum threshold, meaning that their most egregious injuries are treated sequentially until the point of stabilization. Thus, the ultimate goal is to touch base with all patients to

perform a rough initial sorting and some basic care. Through this process, tasks begin to be further defined and a procedure is formulated to engage in the broad-scale triage efforts. Each patient observed and identified updates the system state. This process iterates until the field requires a lower level of cognitive effort. Should a responder interact with an evaluated patient, they will take the branched path after 'seek most emergent patient' to 'Provide follow-up care.' Respondents will switch to the Economical strategy once the cognitive effort requirements are sufficiently low.

Essentially, the rapid strategy requires responders to make 'quick and dirty' classifications with minimal information, thus sacrificing (or costing) accuracy to ensure a rough class of patients are attended to. Another sacrifice is that of data acquisition, i.e., additional pertinent information on each patient. This loss of nuanced accuracy and time means follow-up waves are less immediately meaningful and that some patients are perhaps not prioritized correctly. The lack of data yields a decreased effectiveness of communication of patient status across responders (or 'handoff'), yielding less effective care.

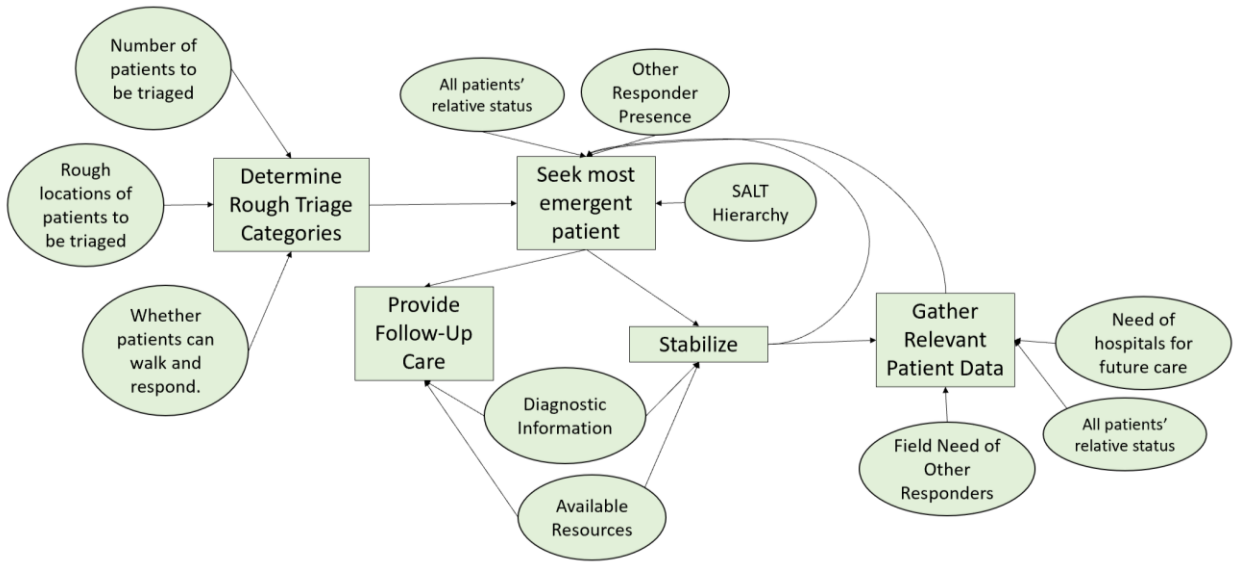

Figure 11: *Information Flow Map for Economical Strategy*

The Economical strategy is a middle ground between spending the minimum time required to stabilize each critical patient and affording more qualitative time with critical and non-critical patients (lower-tiered triage categories like yellow and grey). This allows for a more granular understanding of the field status and for responders to better understand each patient on an individual level. Thus, this information flow map includes new elements associated with data collection. However, there is still a broad patient base in need of care, meaning that responders still cannot gather all relevant information or provide reassurance with their constant presence. Patients receiving an economical strategy may receive more holistic care and care beyond the minimum required for stabilization but not as much as they would in the attentive strategy. Thus,

the ultimate goal has changed from simply making snap judgements about the group as a whole to backfilling relevant information, providing additional but less emergent care, and spending more time with less critical patients. The task definitions change to be more patient based, and the system state is better understood. This process iterates until, again, the field requires a lower level of cognitive effort. Once this effort requirement has been sufficiently lowered, responders will switch to the final strategy: Attentive.

Again, we see costs associated with information (data) collection and accuracy, though these costs are lower with this strategy. The tradeoff, however, is that fewer patients can be 'seen' in the same amount of time. Handoffs are more meaningful and more effective with patient treatments. However, given the nuanced state of patients and the total number of them coupled with a better but still limited accuracy, some patients are not seen as promptly as they should be.

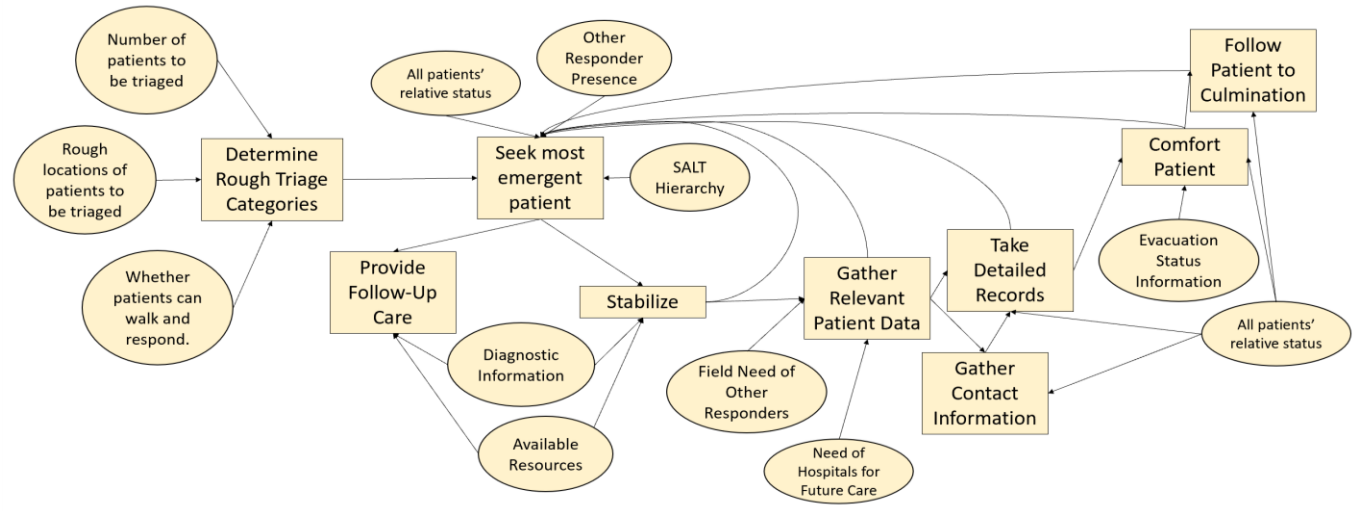

Figure 12: Information Flow Map for Attentive Strategy

The Attentive strategy denotes responders being able to afford individualized, highly qualitative care to patients. This includes spending diagnostic and treatment time with 'green' and 'yellow' patients. This qualitative care can also include staying by a frightened patient's side and will afford more cohesion of care between handoffs. I.e., the patient's field record will be more thoroughly intact and detailed upon reaching the hospital. This affords the most granularity in triage assignment as responders can spend more time with patients to truly assign an appropriate category. Finally, patients may be treated much more holistically and simultaneously, as well as beyond the minimum standard of care. This means that instead of treating each concern in priority order, the patient status is treated entirely (giving medications that handle several concerns simultaneously due to time afforded to consider full patient). Further, patients may get additional, non-emergent care above that available in the economic strategy.

This process will iterate until so few patients remain that the triage categories begin to melt away (assuming no sudden critical changes) as the response experience becomes more typical of standard response scenarios. Eventually, the field will be cleared, and the MCI scenario concluded (as far as EMS is concerned).

This strategy maximizes data collection and accuracy. This data collection is critical as it affords a continuance of care between responders and the hospital. Unfortunately, this is the slowest strategy and affords the least throughput quantity of patient evaluation.

3.5.1 Design Implications – Phase III (STrA)

The selection of and shifting between these strategies is based on the available resources, scene severity, and the perceived cognitive load of the responder. A responder's sense of command over the scene will play a prominent role in determining their strategy selection. Any systems developed to support MCI Triage should, first and foremost, afford the 'wave' process in which one responder arrives and is followed up by a potentially different responder. AR systems should afford these three variable strategies (styles) of response and ensure that the AR systems themselves are quick and efficient (for the rapid strategy) but can also be detailed and nuanced (for the attentive strategy). Ideally, supporting responders enough to reach the attentive strategy is ideal as this strategy affords the greatest continuity of care and the best patient experience. For example, our AR triage tag facilitates the 'rapid' strategy by making the most important facets of the tag (triage category and time last attended, Figure 13) the most salient features on a minimized "card" display. For the most attentive strategy, the full triage tag offers greater fidelity via verbal note tracking with embedded text-to-speech.

Figure 13: The minimalistic 'card' for the triage tag

It is important to note that these strategies are not linear, meaning that responders will not always start with the rapid strategy and work their way to the attentive strategy, where they conclude/stay. Given the severity of a scene and the number of casualties, a responder could very well begin with the attentive strategy. However, if the scene drastically changes, that strategy could switch to the rapid strategy. This means that support tools must be flexible to afford dynamic shifts in strategy. We posit that the bulk of MCI response is spent in the economic strategy. Thus, the goal is to achieve some averaging or bridging of the three strategies, or to facilitate greater data acquisition and accuracy in the rapid strategy and perhaps increase efficiency of patient evaluations for the attentive strategy.

3.6 Phase VI: Social Organizational and Cooperation Analysis (SOCA)

The next phase of the CWA is the social organizational and cooperation analysis (SOCA) and investigates how the responsibilities of the work domain are distributed across responders and how responders communicate [92].

On-the-scene responders are responsible for all facets of the decision ladder outlined in Phase II. However, some elements, including other human parties, can be shared, enhanced, or partially offloaded. For example, other on-the-scene responders can assist in general, immediate domain state awareness. I.e., These other hands-on responders will also be observing the work domain and identifying elements like patient status, remaining resources, etc. Thus, coordination and communication with fellow responders can ensure the 'self' responder has a broader understanding of the work domain.

Furthermore, this means that responders have improved access to the complete pool of the team's experience. Scene-responders and all their communications are then overseen by an incident commander [36], [35], [42]. Incident commanders can offer further support in understanding the work domain's comprehensive status (situation awareness) and the interpretation of the configural relationships among relevant variables and through the provision of decision support. Incident commanders also help coordinate the team members themselves, thus enhancing the team's ability to leverage all relevant expertise. Thus, a degree of cognitive offloading is possible when working effectively with hands-on team members and commanders (responders are allowed to interpret a smaller slice of the work domain as opposed to the work domain in its entirety). Furthermore, leveraging these resources improves a responder's understanding of the entire work domain (elements not directly present to the hands-on responder), thus allowing for a more efficient response.

To better understand the coordination of information and responsibilities among responders and incident commanders, it is essential to understand their communication structures. Similarly to the military, there is a formal, rigid, and centralized hierarchy [93], [92] among emergency responders with clear chains of command. In times outside of MCIs, this steep hierarchy governs the bulk of the responder's daily operations (restocking the ambulances, cleaning the bay, delineating training requirements, etc.). However, much like Rochlin et al. identified with navy aircraft organizational structures [93], complex emergency response operations like MCIs result in a profound departure from this centralized, hierarchical structure.

During MCI triage response, the organization structure of the team shrinks and becomes informal, flat, and distributed as opposed to hierarchical and centralized. This is because events impacting on-the-scene responders are quick and can require action faster than appealing through the chain of command would allow. Thus, when on-the-field, each responder has the authority and responsibility to make judgment calls they would not typically have the authority to make (like whether to evacuate an immediate area, search a

nearby area, etc.). Further, if an incident commander makes a dangerous call, on-the-scene responders have a duty to ignore that command in lieu of a safer course of action. The exception is if an incident commander is ordering a scene-responder to back out of an area due to safety concerns. I.e., scene-responders cannot override reasonable commands intended to preserve their safety.

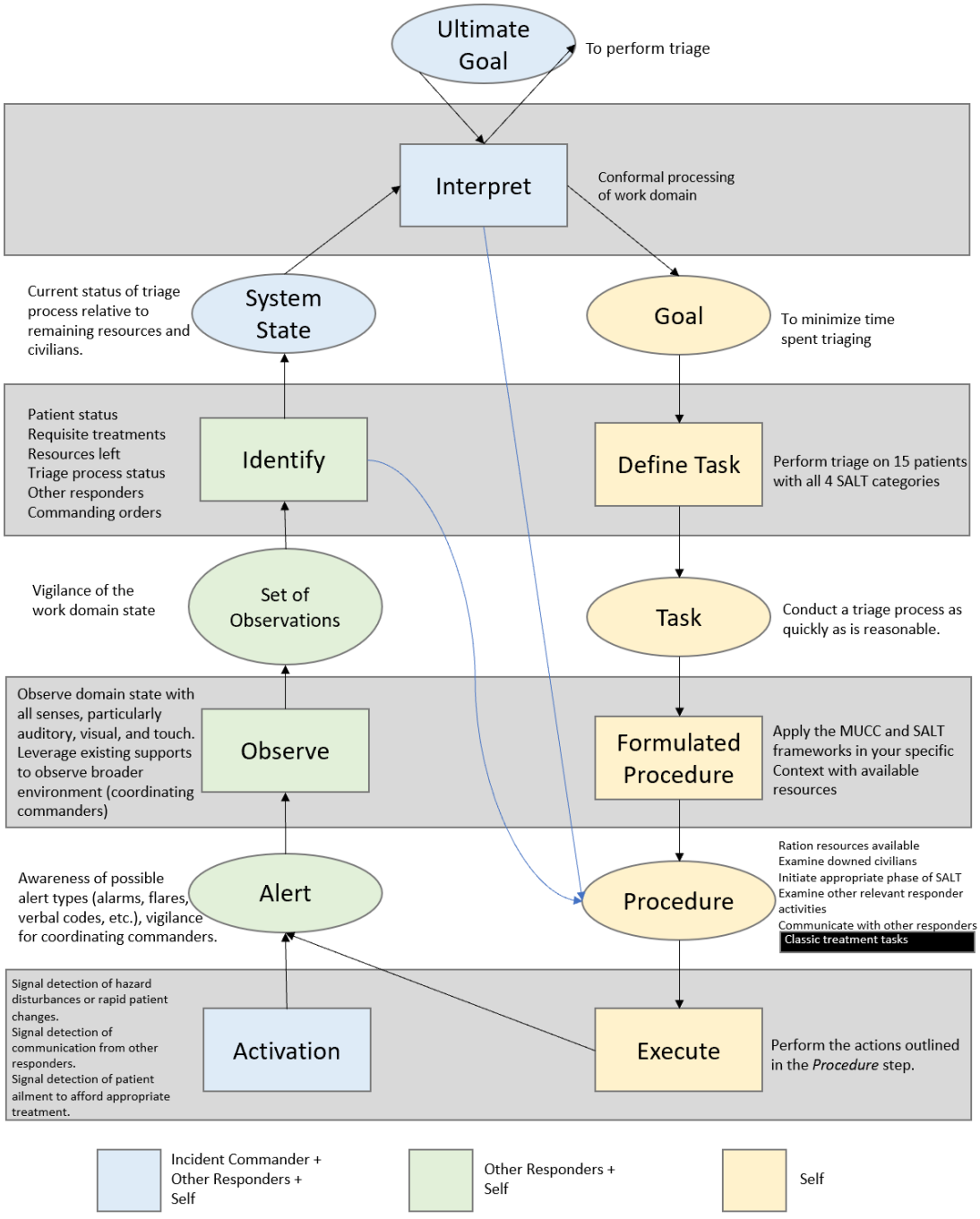

Figure 14: Mapping Social Coordination Among Responder Control Tasks

Figure 15, heavily inspired by Vicente's example [92], shows a breakdown of how incident commanders oversee on-the-scene responders, how these scene-responders interact with each other, and the realms of responsibility among each responder concerning the actual work domain. This is what we meant when we said responders must complete the entire decision ladder individually but can benefit from the support of other responders and commanders; responder A can learn much more about the full crisis by speaking to responder C and (overlapping circles) and even more by engaging commander D (to learn about re sponder B's disconnected area).

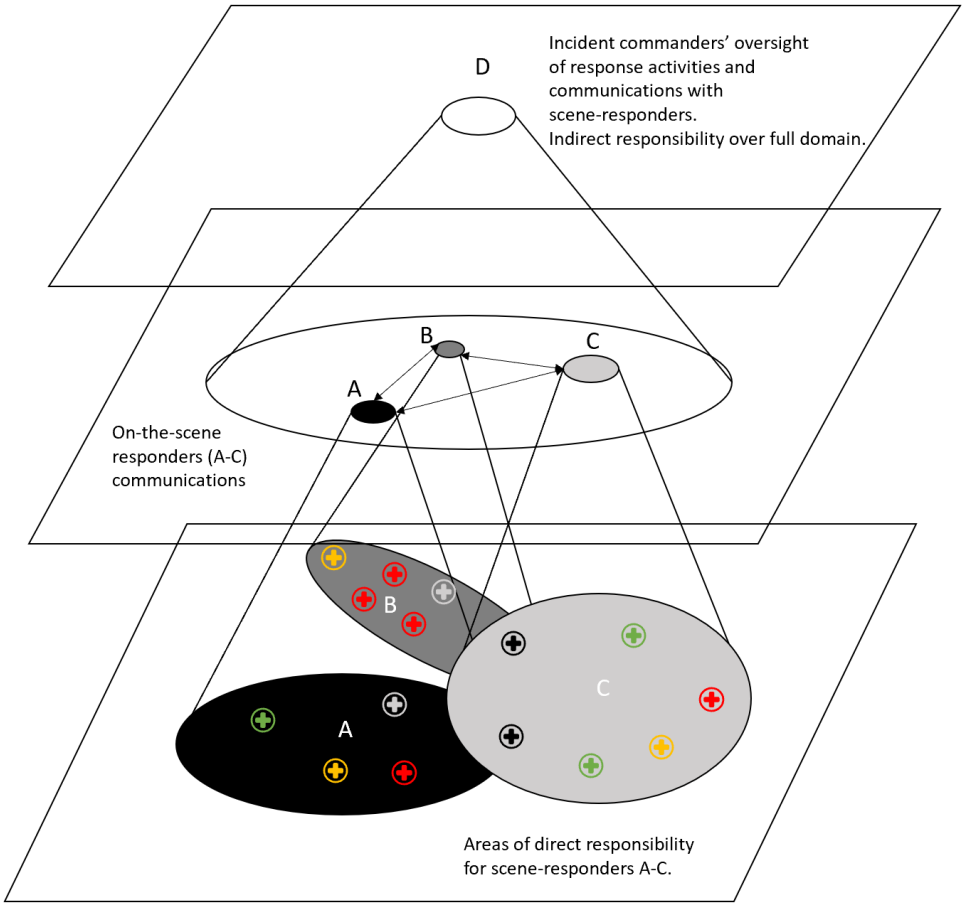

Figure 15: Diagram of Communication Between Responders and an Incident Commander

We noted in an earlier paragraph that the hierarchy shrinks and is, therefore, dynamic. This is because the nuance of lieutenants, sergeants, captains, etc., gets boiled down to only the on-the-scene responders and incident commanders. An interesting consideration not evaluated or captured by our analysis is that the social structure likely shifts in tandem to the shifts in strategies outlined in phase 3. As the work domain

becomes less hectic and more controlled (allowing more attentive response strategies), the hierarchy likely becomes more rigid, nuanced, and centralized.

3.6.1 Design Implications – Phase VI (SOCA)

Future AR systems could bolster the communication between scene-responders and incident commanders to help paint a comprehensive picture of the entire MCI domain by supporting and allowing better efficiency in the communications between responders. In particular, systems that can offload the active process of constant information sharing in any capacity would be beneficial. Related to this, any systems that can support the generation and sharing of configural understandings of the work domain would also be advantageous. More specifically, within MCI triage, this would look like systems that can integrate the information of interdisciplinary inputs across the field (from search and rescue, to EMS, to ambulance, to resource status). Some potential manifestations will be in providing conformal relationship support (perhaps via adaptive AI systems) to help responders interpret the status of the scene. The goal will be to support incident command from afar in their endeavor to maximize the safety of responders on the ground, as well as to facilitate responders themselves to make effective triage decisions. Finally, any systems that intend to offload elements of an individual responder's decision ladder must ensure that doing so does not hinder the integral communication and resultant understanding of the developing crises.

3.7 Phase V: Worker Competency Analysis

The next phase of the CWA is the Worker Competency Analysis (WCA), which investigates what knowledge, rules, and skills workers need to be successful in the work domain [94]. For this purpose, we leveraged Rasmussen's Skills, Rules, and Knowledge (SRK) inventory [95]. Our analysis was heavily inspired by Kilgore's SRK inventory on a rerouting control ask for flight control [74].

Our SRK inventory outlines a responder scanning an MCI scene (associated with all three strategies and within the control task of identifying the system state), determining the rough triage statuses of all patients within their cordoned area of responsibility (CAOR), predicting future patient states if they are tended immediately with current resources available, determining the criticality of patients that might be savable with current resources if not tended immediately, selecting the most critically treatable patient to tend first, selecting a strategy for that process based on the relative triage domain, and finally executing the selected strategy and other classic medical control tasks to stabilize the patient.

An important limitation to note is that our SRK did not include/capture the nuance of patient selection relative to hazards on the field. This is much more nuanced and thus does not have an established, go-to process. I.e., choices made in those circumstances are highly dynamic and case-by-case, thus this facet was

excluded from our analysis. The bottom two rows are melded together because the selection of the strategy and execution will happen interchangeably. I.e., a responder might shift between strategies as they are executing medical care control tasks.

Table 3: SRK Inventory of Scene Assessment and Patient Selection

Information processing steps	**Resultant Knowledge State**	**Skill-Based Behavior (Signal)**	**Rule-Based Behavior (Sign)**	**Knowledge-Based Behavior (Symbol)**
Scan for patient and other responder presence within MCI cordon area of responsibility (CAOR)	Whether there are patients and other responders in CAOR	Monitoring overall triage status within CAOR	Perceive explicit indication that there are patients or responders within CAOR (Patient, e.g., flares, calls for help, waving, etc.) (Responder, e.g., flares, emergency lights, etc.)	Reason, based on presence of patients and responders, that several patients might need to be seen within similar time frames (e.g., cordon tape, missing limbs, poor responder to patient ratio, etc.)
Determine rough triage status of all patients within CAOR not presently being seen by another responder	Whether there are multiple patients with CAOR in need of urgent tending	Perceive patient statuses as dead, expectant, immediate, delayed, or minimal	Use heuristics (EMS training) to determine if patient statuses are stable or deteriorating (e.g., visual and auditory alerts from diagnostic tools, protruding bones, bruising, laceration location, relative blood flow from laceration, exposure to hazardous materials like radiation, etc.)	Reason, based on heuristics (medical knowledge) and informing biological data whether patients appear to be stable or deteriorating (e.g., blood flow in ml from injuries, exact heart rhythm behaviors from EKG, etc.)
Predict future patient states if they are tended immediately with current resources	Whether multiple patients are trending toward a critical state in similar time frames that can be saved with current resources	Perceive time to criticality shift from red (immediate) to grey (expectant)	Use heuristics to estimate multiple patients are trending toward a critical state shift beyond a point of return (red to grey) in similar time frames (e.g., visual and auditory alerts from diagnostic tools, protruding bones, bruising, laceration location, relative blood flow from laceration, exposure to hazardous materials like radiation, etc.)	Calculate blood flow rate, breathing rate, heart rhythm, rough time at which patient will deteriorate beyond point of return for field interventions (e.g., unrealistic for human in a full mathematical form. Not enough time, data, and too much variance between people)
Determine the criticality of treatable patient injuries if not immediately tended	The degree of severity for multiple patients if not tended immediately	Perceive whether there are patients with more severe outcomes if not tended immediately relative to other patients	Use heuristics to estimate priority order of patients based on their relative dynamic statuses and deterioration rates (e.g., visual and auditory alerts from	Using comparative calculations of blood flow rate, breathing rate, heart rhythm, and other heuristics to determine priority order of patients based on dynamic patient status relative to other dynamic patient statuses (unrealistic for human in full

			diagnostic tools, protruding bones, bruising, laceration location, relative blood flow from laceration, exposure to hazardous materials like radiation, etc.)	mathematical form. Not enough time, data, and too much variance between people for prediction validity of each person)
Select most critical, treatable patient to tend	Which patient to tend next based on who has most critical outcomes if not tended while they are still savable with current resources	Directly perceive which patient is in most need of care	Use heuristics to identify the patient with the most critical outcomes if not tended while still savable with current resources (e.g., visual and auditory alerts from diagnostic tools, protruding bones, bruising, laceration location, relative blood flow from laceration, exposure to hazardous materials like radiation, etc.)	Using comparative calculations of blood flow rate, breathing rate, heart rhythm, and other heuristics to determine priority order of patients based on dynamic patient status relative to other dynamic patient statuses based on current resources (unrealistic for human in full mathematical form. Not enough time, data, and too much variance between people for prediction validity of each person)
Select a specific strategy for tending to selected patient (rapid, economical, or attentive)	Responder's awareness of the triage domain after selected patient has been tended	Respond automatically to perception of patient criticality relative to field and only stabilize patient	Classify patient more holistically based on rough hierarchy of their injuries and select appropriate classic medical treatment control task for most egregious injury	Classify patient holistically with nuances and treat entire human system, considering all injuries in treatment administration. (instead of treating sequentially, treating simultaneously where possible)
Execute specific strategy for stabilization execution	Awareness of shift in triage domain (one less untended patient), plus time spent with that patient and the realization of earlier predicted patient trajectories			

3.7.1 Design Implications – Phase V (WCA)

Skill-based behavior (SBB) is not feasible for this work domain as the bulk of sample problem space is based on the dynamic interpretation of rule and knowledge-based domain elements. There is potential for SBB in the actual treatment administration process, just not in determining patient statuses. To support rule-based behavior (RBB), better field diagnostic tools could be created to make it easier for responders to determine the signs associated with patient status. Example tools are pulse oximeters, blood pressure cuffs, glucometers, etc. RBB should be the foci of future system considerations, particularly as knowledge-based behavior (KBB) is too difficult for human responders to do. RBB's signs share many similarities with KBB's symbols. An example of this is that blood flow is a consideration of both RBB and KBB. However, RBB is a rough estimation of blood flow and its relationship to where it is coming from on the body. Alternatively, KBB is a detailed understanding of the exact ml blood loss over time relative to the person's total blood volume, rate of loss, and the amount already lost, compared to standard expectations provided via medical knowledge heuristics. Thus, KBB is unrealistic to consistently afford in a field setting due to

the case-specific nature of KBB scenarios (E.g., planning a coordinated effort to save a specific individual near a hazard requires some level of KBB).

Regarding competencies, responders need a working understanding of several medical heuristics related to the human body to appropriately and successfully determine patient statuses quickly. Further, responders need training in MCI triage to adequately participate in the shared, categorical priority assignments of patients. Moreover, the SRK helps us understand the information requirements of responders in the MCI triage domain. Finally, responders required an evolving, nuanced understanding of the entire CAOR, including detailed information about patient statuses and available resources so that responders can make informed choices about which patients to stabilize next. Systems that can support knowledge requirements in these categories would be immensely helpful, followed by systems that could go so far as to make those predictions for the responders. Any informational displays should aim to elicit RBB by using signs in the interface. Basic information that should be available on those displays based on our WDA, CTA, and Strategies Analysis include resource availability and status, scene safety and hazards, patient status (broadly) and individually, locations of each responder, information provided by incident command and dispatch, as well as any conformal relationships (ETA of an ambulance with x number of bandages at y geolocation relative to z patients).

3.8 Conclusion

Our analysis gathers a firmer understanding of MCI Triage, a micro-component of MCI broadly. Through this analysis, design considerations for AR have emerged that can support effective AR tool development aid responders in treating patients during the 'golden hour' of trauma. Each phase answers a critical question about the work domain:

- Phase 1 – Why? (Abstraction Hierarchy Decomposition)
- Phase 2 – What? (Decision Ladder)
- Phase 3 – How? (Information Flow Map)
- Phase 4 – By whom? (Mapping to Decision Ladder)
- Phase 5 – By what means? (SRK Inventory)

Information gleaned from Phase 1 propagated and informed the analysis at Phase 2, and so on until culmination in Phase 5. Ultimately, we have learned that despite the inherent disturbance-laden features of MCI triage, there are opportunities for computer-based systems to support emergency responders.

The primary takeaway is how critical responder situation awareness is for responders to successfully meet the functional purposes of the work domain. The element of situation awareness consistently propagated as a core consideration at every phase of the analysis and should thus be a core consideration of any future design work. This finding is consistent with the trend of current literature in the emergency response

space, putting a premium on situation awareness [3], [11], [36], [37], communication among responders in relation to situation awareness [19], [32], as well as the call for greater situation awareness research in the field of human-computer interaction generally [26]. A core component of that situation awareness is that of the communication happening among responders during MCI response.

The benefit of conducting this analysis, in addition to observing research trends and designing in the MCI triage space, is that the a usability evaluation and research trends alone do not give us a nuanced understanding of the pervasiveness of situation awareness within the nuances of inter-responder communication, or a deep understanding of the informational and cognitive requirements to afford situation awareness. This evaluation provides a means to ensure that future AR designs unrelated to situation awareness do not indirectly hinder it, and we now better understand the informing elements and the cognitive requirements in understanding true AR facilitation of MCI triage. Future work should expand upon this analysis and include elements not included not only in the domain of triage but MCIs more broadly.

4 Study 3 – User-Centered and Situated Evaluation of an AR Virtual Triage Tag Comparing Virtual and Real Patients: Usability, Perceptions, and Performance

4.1 Introduction

Mass casualty incidents (MCIs) often create more casualties than typical resources can manage [1]. This leads to a breakdown of routine procedures and thus demands extraordinary assistance [1] to handle poor management, planning, communication, and resource management [2]–[4]. These issues are exacerbated by the vast range of possible disasters and the relative rarity of MCIs. Worded differently, MCIs are chaotic, dangerous, and infrequent events that upend standard procedures like triage and exhaust limited resource pools. Triage is a core part of early MCI medical response and seeks to "… apply available medical resources most effectively" [96] during the 'golden-hour' of triage where treatment efficacy is highest [7]. Ergo, triage requires effective coordination of responders, resources, and scene handling. Earlier work identified augmented reality (AR) as having the capacity to provide exceptional support [25] to ERs. To this end, we identified AR's use cases in ER and augmented the existing SALT (sort, assess, lifesaving interventions, treatment/transport) triage process with an AR virtual triage tag (VTT). This was done via a user-centered design and evaluation process with emergency medical responders (chapter 2, [25]), and we evaluated the complex work domain (chapter 3). Of the identified use cases, we created the Augmented Reality Triage Tool Suite (ARTTS). ARTTS was designed to facilitate:

1. **Setting the stage** by setting up evacuation points, finding patients, stabilizing the scene, etc. Tasks that aid in scene processing with both quick data collection and a decision-tree system,
2. **Communication** by providing a means to record, update, and display real-time data in a way that can be communicated to other responders in order to support triage, and
3. **Supporting decision-making** by lessening cognitive loads associated with remembering the triage process and incident tracking, jolting stunned responders back to the triage task, and by providing situation awareness support of patients.

Our prior evaluation determined that the virtual triage tag (VTT) had average usability in a lab setting with representative user tasks [25]. For actual implementation, more context-sensitive studies are needed. Due to the high-risk nature of MCI triage for both responders and patients, it is impractical to jump straight to actual field studies. High risk design environments require a smoother continuum for testing, which is where simulations find a foothold. Moreover, this high-risk work environment also begets a demand for simulation-based training. The relationship between the demand for training simulations and the opportunities for simulation-based evaluation go hand in hand, and this work seeks to leverage that relationship.

Many emergency medical responders (often the first on scene [82]) are not prepared for MCIs, with "adequate preparedness [being] long, difficult, and expensive [to obtain]," [81]. Relatively low MCI frequency compared to standard EMS calls means 1) limited general experience for emergency responders (ERs) about successful coordination and triage under difficult MCI conditions and 2) even for responders privileged enough to have specialized training, it is unclear "… whether knowledge and skills obtained through education and training will be retained" for application when next needed [81]. Wealthier and larger cities can offset this some through happenstance (a larger frequency of MCIs due to a larger population), and through funding mechanisms unavailable to many smaller and rural municipalities. Despite the fact that rural responders serve ~75% of the nation and 49 million people, rural ERs have fewer certifications, experience, and financial resources broadly [81], [82]. This means that most ERs are already less familiar with MCI triage than necessary, and this would undoubtedly extend to the usage of any future AR-based cognitive support. Thus, there is a profound need for simulation-based training or re-ups for emergency medical providers, especially in "challenging rural and remote settings," [97] (though urban ERs would benefit too). The existing 'gold standard' strategy for training in MCIs is to send ERs to specialized training facilities scattered across the United States like TEEX's (Texas A&M Engineering Extension Service) Disaster City. Unfortunately, this is too cost-prohibitive for smaller EMS localities. Further, small response units cannot withstand the loss of their responders for extended interstate travel. This alone points to an interesting intersection of a need for station-based training, a niche AR can fill, and AR's established identification for this ER use-case. Interestingly, facilities like TEEX also desire virtual simulation-based options in order to minimize the time ERs must physically be on site and to prime them for maximum success upon arrival, dovetailing with XR based virtual simulations in emergency response blossoming as a research area [43], [98]–[102] due to AR's ability to furnish realistic and immersive environments [98]. Compellingly, there is an unanticipated relationship between rural ER stations and academic researchers regarding limited resources. A UCD context-intensive evaluation can be both expensive and time prohibitive for an academic research cycle that demands rapid publications. Thus, this need for training simulations can be combined with our evaluative need, pooling our resources to create both a training simulation and testbed for evaluating new tools for MCI response, like the VTT. Note that this chapter will not focus on the training outcomes of the simulations described herein as our colleagues have another pending paper outlining those findings (*under review redacted).* Instead, this chapter will focus on the methodological contribution of the simulation designs and how they were leveraged to evaluate the VTT in a training context. While there are virtual reality simulations for MCI training [103], [104], to our knowledge this is the only augmented reality simulation.

We partnered with TEEX's Disaster City to create and host a user study evaluation of a virtual simulation and a 'real' (patient actor) simulation. More specifically, we sought to simultaneously: 1) Understand

how the virtual patient simulation compared to the real patient actor simulation, and 2) evaluate the usability of the VTT via quantitative measures of performance and qualitative reports from emergency responders. This work outlines two simulation strategies to not only offer a more robust evaluation of the VTT, but also to provide some grounding for AR simulation viability in emergency response - as HCI developers continue to evaluate tools with emergency responders, simulations will be key to meaningful evaluation. We offer a factor analysis of what responders consider when evaluating a new technology, as well as some design considerations for future AR development in the context of those factors.

4.2 Background

4.2.1 The SALT Triage Process

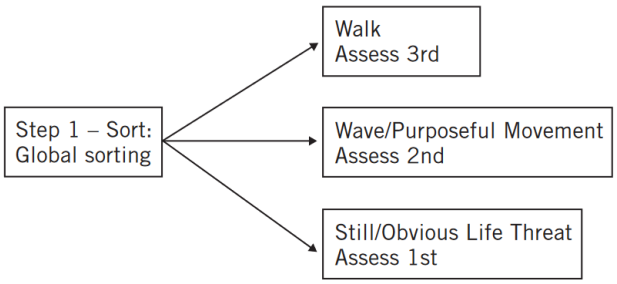

Figure 16: Phase 1 of the SALT flowchart [37].

To understand the foundation of the trainings and test-bed simulations for the user studies, let us first define the SALT triage process. The Sort, Assess, Lifesaving Intervention, Treatment/Transport (SALT) framework [37] is a best practice strategy for triaging patients in an MCI. SALT is a 'wave' based flow chart with a minimum of two waves. Step (or wave) one is to globally sort the patients. This provides a rough-cut prioritization of patients on the field based on their ability to respond to a command and move toward a responder. A responder will shout to the field something to the effect of "If you can hear me and move, make your way towards me. If you cannot move but you can hear me, wave." If a patient is still, unresponsive, confused, or has other obvious life threats, they are assessed first. If patients cannot move but understand the responder and wave, they are assessed second. If a patient can understand the responder and can walk or otherwise make their way toward a responder, they are assessed last.

Step (or wave) two is to then provide individual assessments. With these rough prioritizations, a responder makes their way to a patient in the first priority category and assesses them for several attributes that define an ultimate triage category. A *triage category* denotes both an overall assessment of the patient's current medical condition and determines a finer priority ordering for resources and evacuation [25], [37].

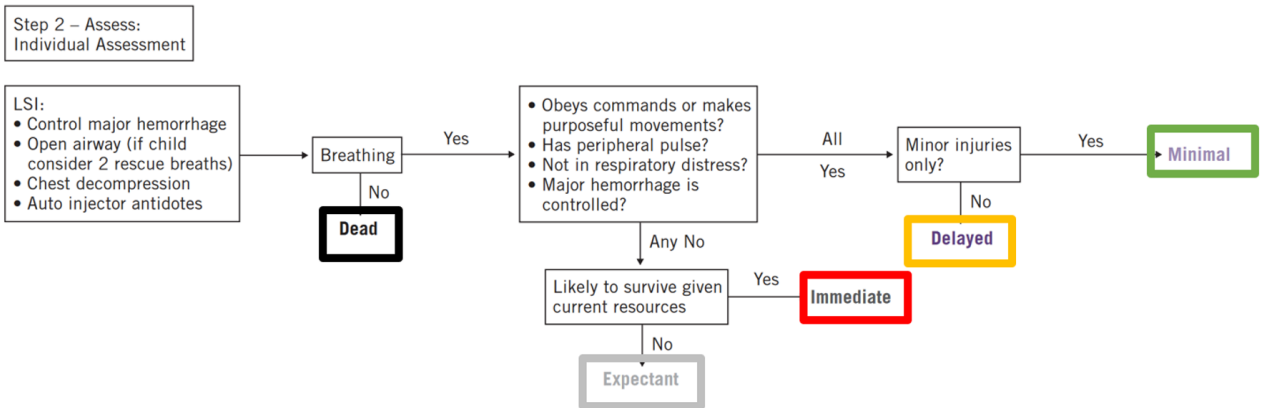

Figure 17: Phase 2 of the SALT flowchart. [37]

There are five categories:

- Black: deceased patients (last priority),
- Grey: patients expected to expire or for whom current lifesaving resources are not sufficient or unlikely to save (third priority),
- Red: patients needing immediate care in critical condition that can potentially be saved with a combination of field resources, interventions, and medical intervention off of the field (first priority),
- Yellow: Patients who can withstand delayed care after receiving stabilizing care in the field but will still need follow-up medical care (second priority), and
- Green: patients experiencing minimal to no injuries (fourth priority).

Once these initial and required two waves are completed, responders will continue to rotate (or complete waves) between patients based on finer and finer nuanced needs of the individual patient.

4.2.1 Updates to the Virtual Triage Tag

The Virtual Triage Tag is the core AR triage tool leveraged by our user study. It is based on the tag defined and outlined in Chapter 2, and updated based on findings in Chapters 2 and 3. In true iterative UCD spirit in tandem with our CWA, we have leveraged responder feedback from a previous study [25] to improve upon the VTT. Namely, we

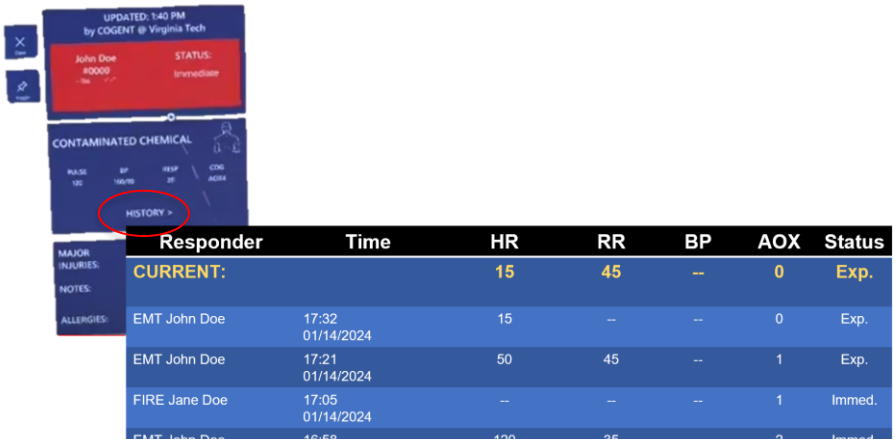

Figure 18: The new history button and what the history button output looks like

Figure 19: The minimalistic 'card' for the triage tag

have added an additional "History" button that tracks all reported data and provides it in a tabular, time-stamped format for ease of responders in order to facilitate more effective situation awareness and conformal understandings of individual patients. The tool suite still leverages a HoloLens 2 and affords poke, raycast, and voice interactions with the VTT. We also added the 'god's eye' view of the triage field that allows responders to look over the field and see 'summary' cards. These summary cards denote the triage category of a patient (if assigned) and how many minutes it has been since the patient was last attended to. Finally, we have added the ability to 'spray paint' to mark the scene, though this new feature is not evaluated by this study.

4.3 Methods

4.3.1 Participants

This study had 13 participants. Two are female and eleven are male (15/85%), with an average age 36.23 (21 to 60). All participants identified as White/Caucasian. Ten participants have EMT-Basic or EMT-Paramedic level certifications, two have experience in emergency response broadly, and one had CPR and First Aid training. Ten have had MCI triage training in the past and six have had three or less hours of AR/VR experience while two have had 5 or more. Regarding their representativeness, nationally, ~85% of responders are White/Caucasian while only about ~2% are Black/African American, and ~76% of responders are male with an average age of 38 [105].

4.3.2 Study Design

This study leverages a within-subjects mixed methods design and is separated into 4 phases containing 2 conditions, first defined and then further explained below. The independent variable is 'condition/simulation' between the virtual patient (1V) condition and the 'real' patient actor (2R) condition The study phases are: 1) Training in Triage, AR, and AR triage, 2) Virtual Patient Simulation [Condition 1], 3) 'Real' Patient Actor Simulation [Condition 2], 4) Semi-structured interview debriefs. The simulations were designed in collaboration with TEEX and a subject matter expert with 30+ years of experience as a responder and a trainer of other responders in MCI triage.

We captured the following dependent variables: time (in seconds) spent with each 'patient', whether the responder triaged that patient correctly, how many buttons responders pressed during each patient session, and the direction of the triage error (if one was made). Then, the study concluded with semi-structured interviews. We also captured motivation to participate in the study and bias scores via the Implicit Association Task (IAT).

Both simulations required formally defined 'patients.' With the aid of a primary SME, we defined the characteristics of twenty total 'prospective' patients to serve as study stimuli in a "Master Case List." Given the relative ease of identifying dead vs. healthy patients, we represented the categories with variable weightings. More specifically, the pool contained three patients in the black/dead category, four for grey/expectant, five for red/immediate, five for yellow/delayed, and three for green/minimal. The MCI scenario given to the ER participants was that a nearby fertilizer plant exploded. Example category details are as follows:

- Black (deceased): Patient 004 is suffering from a tertiary blast injury form a fertilizer plant explosion. No pulse and not breathing.
- Grey (expectant): Patient 006 is suffering from a secondary blast injury from the plant explosion. Respiration rate is 8 breaths/minute, they do not obey commands, their pulse is 35 beats/minute, they are in respiratory distress, major hemorrhage has been controlled or was not present, they cannot speak, their BP is 170/110, AOX2, GCS 3.
- Red (immediate): Patient 011 is sitting up and suffering from a primary blast injury. HR:130, RR: 12, BP: 90/62, AOX4. Patient complains of abdominal pain and tenderness when palpated.
- Yellow (delayed): Patient 030 is suffering from a quaternary injury from the explosion but is able to stand. HR: 118, RR 28, BP: 120/80, Pulse – strong/regular SpO2 90%. Patient is wheezing.
- Green (minimal): Patient 040 has a quaternary injury from the explosion and is standing. HR: 99, RR:22, BP: 126/80, Pulse – strong/regular, AOX4 and can walk to ER.

The simulations were programmed to capture response time (based on the moment a responder opened a VTT until they closed it), button presses (how many buttons were pressed on the triage tag), and the triage category input by responders. The triage category data entered by responders was run against an internal master key to provide an additional data output of whether the responder's answer was correct and, if not, what had been entered vs. what should have been entered.

*4.3.2.1 **PHASE 1:** Pre-Simulation Training Procedure*

All participants were given an extensive, 4-hour computer-based training developed by TEEX on both MCI triage and the AR VTT (which is evaluated and outlined in our colleague's aforementioned *redacted* working paper). This training included several videos [106] and text-based descriptions of the AR VTT and its interactions. For basic AR interaction training, participants were given an AR familiarization experience for ~30 minutes that taught poking, raycasting, object movement, and scrolling. Low performers were given additional exposure to these AR tasks. More information on this adaptive component of the AR training protocol was published by our colleagues, S. Vyas et. al., (2023) [107]. Participants were then given ~30 minutes of experience using the VTT in AR before final evaluation.

*4.3.2.1 **PHASE 2:** Virtual Patient Simulation (User study Condition 1V)*

Once participants were familiar with fundamental AR interactions, participants engaged with the virtual patient simulation (condition 1). This task entailed triaging five virtual patients within ten minutes. We opted to leverage five patients for the virtual simulation due to likely space constraints in a given rural fire or EMS station. Details for these five patients were provided by the aforementioned "Master Case List" pool, with one patient from each category represented. The five virtual patients had variable animations to communicate their injury status, like a patient hopping on one leg due to a fractured ankle.

Figure 20: What the researcher sees when a participant engages with the virtual simulation vs. what the participant sees.

ERs were able to interact with virtual patients in several ways. In order to attain blood pressure, participants put their palm on the bicep of the virtual patient, which would create a 'read out.' For pulse, ERs put two fingers on the wrist of the virtual patient for another read out of the relevant numbers. For respiration rate, participants could put their head close to the chest of the virtual patient and hear respiration. For alertness and orientation, participants could ask the patients to "wave" and to "show me where it hurts." Patients who could would wave and gesture to their injuries.

*4.3.2.2 **PHASE 3:** 'Real' (Patient Actor) Simulation (Condition 2R)*

Participants then transitioned to the 'real' simulation (condition 2) with patient actors (recruited via a local acting troupe). The 'real' patient actor simulation involved triaging twenty patients in ten minutes. Of note, the real simulation was done simultaneously with another responder as a team task. This yielded a standard organic strategy among responders of breaking the patients up so there were total of ten patients per responder. This team task was also used to gather respondents' perceptions of the shared VTT workspace with a colleague as when a VTT is closed, a summary 'card' appears, indicating the overall status and minutes since a patient was last attended to.

Figure 21: A responder triaging a 'real' patient shadowed by a trainer.

Figure 22: The summary 'cards' for the information placed on a VTT.

Patient actors were assigned a patient number and given the associated information for that patient. This slip of paper included BP, RR, HR, AOX status, and a description of the patient's injuries. As patient actors are healthy, a direct interaction where responders feel for a pulse is impossible. Thus, each ER participant was accompanied by a training lead. This lead person was an ER employed by TEEX. When an ER participant would reach for the patient's pulse, the training lead would tell the responder what the pulse was. A responder engaging with a 'real' patient and shadowed by a lead trainer is featured in Figure 21.

Patients were laid out in an attempt to mix the triage categories (see Figure 23) thoroughly. Participants started from the bottom of the floorplan and worked their way to the top. One responder usually went through the left doorway near the bottom, and the other typically went through the right. We deliberately did not equally represent each triage category. E.g.: Green (minimal) appeared 3 times, yellow (delayed) 5 times, red (immediate) 5 times, grey (expectant) 4 times, and black (deceased) 3 times. This is because we predict that dead/green patients will be the easiest to determine and that there will be more opportunities for errors in the other categories.

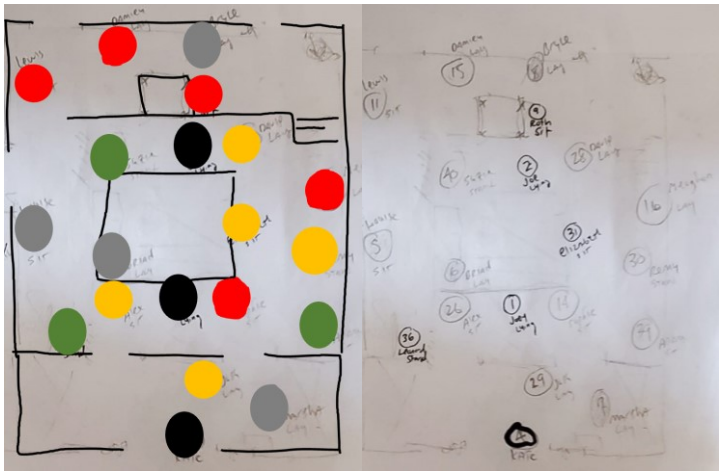

Figure 23: The floorplan of the 'disaster theater' outlining both the patient ID's and the triage category for each patient.

*4.3.2.3 **PHASE 4**: Semi-Structured Interview Debriefs*

Upon completion of the two simulations, the responders were interviewed in the same teams of two using the following questions:

- What did you think of the AR technology?
- Would you use this AR technology in the field?
    - Why or why not?
- What did you think of the triage tag?
- What did you think of the patient filtering?
- What did you think of the hazard marking?
- What did you think of the virtual patients (from this morning)?
- What did you think of the interactions available to you?
- What did you think of the collaborative functions in the system? (Being able to see each other's progress)
- What did you think of poke interactions?
- What did you think of raycast interactions?
- Are there any limitations or concerns you have about this system being deployed in the field?
    - If so, what are they?
- Are there any unique strengths you can see about using this system in the field?

    - If so, what are they?
- What did you like about the AR system?
- What didn't you like about the AR system?
- Would you change anything about the system?
    - Yes/no, why or why not?
    - Is there anything additional you think should be added to the system?
    - Is there anything you think should be removed from the system?
- How would you like to interact with AR interfaces? (I.e., other than poke and ray, are there other ways you might like to interact with the interfaces)
- What would you do to improve the collaboration between yourself and other responders with this system?
- Surprises in learning: Today you spent the morning doing training. What did you think of the training? Did you feel adequately prepared or was there anything you wish you'd learned?

4.4 Results

4.4.1 Variables Defined

These are the independent and dependent variables considered by this paper:

IV	**DV**
Simulation Type	Time
	P-score
	Error Direction
	Button presses
	Overall Accuracy

Though we consider both time (T) and accuracy (A) as their own separate variables, we also wanted to evaluate the relationship between them qualitatively. To make it easier to interpret the data, we created a P-Score with the following formula:

$$\left(\frac{T_{max} - T_x}{T_{max} - T_{min}}\right) + A(1\ or\ 0)$$

For a triage context, less time spent is preferred. Thus, we take the maximum time for the entire set, subtract the T under observation, and divide that by the maximum time minus the minimum time for the set. This means the higher the T score, the 'faster' the participant was. From here, we add 1 or 0 based on the triage accuracy for this observation. As we are looking at overall triage assignment, the accuracy is a binary (1 = correct, 0 = incorrect). Sometimes formulae for this purpose include a W for weighting time or accuracy as more important than the other. For our purposes, we treated each with equal weights.

Scores <1 = incorrect,
<x.5 = slow
>1 = correct
>x.5 = fast
1 = fastest but incorrect
2 = fastest and correct

Figure 24: Interpreting P-Scores

The P-score thus has two components to interpret (see Figure 24). What comes before the decimal indicates accuracy, and what comes after indicates time. So scores <1 are incorrect answers and scores >1 are correct. After the decimal, a score <x.5 means a slower response time, whereas a score >.5 means a faster response time. Scores of exactly 1.0 mean that this was the fastest trial in the entire data set but the responder was incorrect. Scores of exactly 2.0 mean that this was the fastest trial in the entire data set and the responder was correct. As we are analyzing a binary "correct" or "incorrect" on each triaging instance, this p-score affords more analysis strategies and eases interpretation of the conformal relationship between time and accuracy against other variables like error direction.

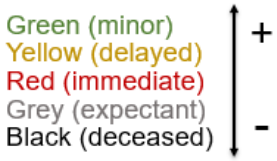

Figure 25: The scale used to determine error direction.

For understanding the nuance of responder accuracy, we also captured the error direction (see Figure 25). Error direction scores range from -4 to +4. A negative score means the responder assigned the patient a more severe triage category than they should have, while a positive score means the responder assigned a less severe triage category than they should have. A score of 0 means the responder was correct. Thus, if a responder triaged a 'Red' patient as green, their error score is +2. If a responder triaged a 'Yellow' patient as grey, their error score is -2. This score helps us understand how far off responders were.

4.4.2 Outliers and Normality

Our analysis began with exploring potential outliers by condition in error direction, time, overall accuracy rate per participant, button presses, and p-score using a Quartile plot and a Huber M-Estimation set to 4 to balance efficiency and robustness. We found and removed one outlier in time for the condition 2R (the 'real' patient actor simulation) of 257 seconds, and one outlier in button presses for condition 1V (the 'virtual' simulation) of 18 button presses.

From here, we explored the normality distribution of error direction, time, overall accuracy rate per participant, button presses, and p-score. We found each of these variables to be normally distributed as a skewness between of -2 to +2 and kurtosis between -7 to +7 means the data is considered to be normal [108] (please see Table 4 for exact values).

Table 4 Skewness and Kurtosis to Ascertain Normality

	Condition 1 'Virtual' Skewness	Condition 1 'Virtual' Kurtosis	Condition 2 'Real' Skewness	Condition 2 'Real' Kurtosis
Time	1	1.585	0.685	0.181
Overall Accuracy	0.358	-1.534	0.24	-0.098
Button Presses	0.876	-0.466	0.524	-0.613
Error Direction	-0.453	-1.253	0.515	-0.613

Performance (P) Score	0.606	0.851	-0.318	-1.446

4.4.3 Missing Data Considerations

This study took place over the course of a full 8-hour day with several moving components and other intellectual contributions not covered by this work. Participants were handed off between teams of researchers for different studies depending on the time of day. Unfortunately, four participants were not given an opportunity to engage with the virtual simulation due to being behind from other studies that day. Fortunately, this means that the missing data is missing completely at random, e.g., the pattern of missingness is not related to the missing values themselves or other observed variables.

We considered two strategies for missing data analysis: 1) Complete Case Analysis and 2) Imputation.

1) The complete case analysis approach involves analyzing only participants who completed both conditions. This gives us the ability to run ANOVA and MANOVA analyses without violating their inherent assumptions. We use this strategy when we are explicitly comparing the two conditions.
2) We considered using imputation to estimate the missing values based on the observed data (via the mode or median of the data). This would allow us to maintain all participants and potentially increase our statistical power. However, our missing data exceeds the 5-10% tolerable amount.

When we are not directly comparing 1V and 2R to each other, we use the full 13 participants available to us. To assess how significantly this complete case analysis may change based on the lower N in 2R, we did a comparative analysis between 2R with 9 participants and with all 13 via ANOVAs on time, overall accuracy, button presses, error direction, and p-score comparing the two data sets for 2R There were no statistically significant differences between the complete case data (9) or the full information data (13). Thus, we proceed with the full information dataset when only looking within 2R to improve degrees of freedom and we do not have cause to worry that the diminished 2R dataset for comparing the conditions will be significantly impacted. These missing data and the complete-case consideration can impact degrees of freedom between tests.

4.4.4 Exploring Performance Within and Between Conditions

For condition 1V, the average number of patients triaged during the allocated time was 3.3. Across all participants, 33 virtual patients were triaged. However, in condition 2R, an average of 7.23 patient actors were triaged, with a total of 94 triaged by all participants.

A MANOVA was conducted to examine the effect of simulation type (Virtual condition 1 vs. 'Real' condition 2) on the dependent variables of time, error direction, p-score, overall accuracy, and button presses. The F test indicated a highly significant $F(4, 101) = 148.867$, $p < .0001$ effect of simulation type.

This effect remained significant after adjusting for sphericity violations (Unadjusted: F(4,416)=153.461,p<.0001; G-G: F(1.0113,105.17)=153.461; H-F: F(1.0214,106.23)=153.461, indicating a robust and reliable effect of condition on the dependent variables. From here, we completed univariate ANOVAs to get a closer look at this significant effect (please see Table 5). Note that condition had a significant impact on participant accuracy, button presses, and the types of errors they made. More specifically, the 2R condition was faster, more accurate, with less button presses. When participants were correct in their answer, they were faster in the 2R condition than 1V. When participants were incorrect, they pressed more buttons in the 2R condition. Condition 1V showed participants tending to round down to more severe triage statuses while 2R had a small effect relative to 1V.

Table 5 ANOVA Comparisons of the DVs by Condition

	Condition 1 'Virtual'	Condition 2 'Real'	F	P
Avg. Overall Accuracy	12.12%	61.05%	(1, 126) = 214.6846	<0.0001*
Performance Score (P-score)	0.7457	1.3931	(1, 126) = 36.3510	<0.0001*
Avg. Time	78.9688	58.2316	(1, 125) = 4.1814	0.0430
Avg. Correct Time	85	45.4878	(1, 60) = 4.6404	<0.05*
Avg. Incorrect Time	77.7143	71.6486	(1, 63) = .1885	0.6656
Avg. Button Presses	3.5937	6.9684	(1, 125) = 14.2063	0.0003*
Avg. Correct Button Presses	5.5	5.7317	(1, 60) = 0.0176	0.8951
Avg. Incorrect Button Presses	3.21	7.73	(1, 63) = 19.3142	<0.0001*
Avg. Error Direction	-0.64286	0.063158	(1, 121) = 8.6927	0.0038*

We then analyzed whether there was a relationship between the time spent with a patient and the accuracy of the triage assignment (see Table 6). There was no relationship between time and accuracy in condition 1V; $F(1, 30) = 0.1533$, $p > .05$. However, there was a significant difference for condition 2R; $F(1, 93) = 4.4598$, $p < .05$, where faster times were associated with higher performance. There was no significant relationship of time on the type of error made OR button presses on accuracy for either condition. Unsurprisingly, there was a significant relationship between time and button presses in both conditions. For condition 1V, each button press is associated with an additional 8.288 seconds, $F(1, 30) = 22.6129$, $p < 0.0001$. For condition 2R, each button press is associated with an additional 7.018 seconds 2R, $F(1, 93) = 63.5613$, $p < 0.001$. We did not find any significant within-participant differences between conditions and there was no relationship between time, button presses, accuracy, p-score, or error direction in 1V yielding a predictable effect in 2R. E.g., higher performers in one condition were not necessarily higher performers in another condition.

Table 6 Comparing Accuracy to Mean Time and Button Presses

	Correct	Incorrect	F	P
Condition 1V Time	85 sec	77.7143 sec	(1, 30) = .1533	0.6981
Condition 2R Time	45.4878 sec	71.6486 sec	(1, 93) = 4.4598	0.0374*
Condition 1V Button Presses	5.5	3.21	(1, 30) = 2.4103	0.131
Condition 2R Button Presses	5.7317	7.73	(1, 93) =1.6928	0.1965

4.4.5 Semi-Structured Interview Analysis

Responder semi-structured interviews were done in the same teamed pairs as condition 2R, giving us five paired interviews lasting approximately 30 minutes each. To parse the interview data, a content analysis was done with four coders distilling a total of 359 comments from the interview transcripts for data coding. Coders were provided 'identities' to code with a 1 (this concept is present in the statement) or 0 (this concept was not present in the statement). The identities were: Logistics, Privacy, Security, Server, Data Access, Expense, Trust, Communication, Usability, Immersivity, Speed, Efficiency, Collaboration, Flexibility, Adaptability, Reliability, Interaction, and Other. Two coders added the additional identity of Learning/Training and one coder added the additional identity of Information. Coders were also asked to indicate the valence of the statement as positive, negative, or neutral. This means that a responder could say "… The AR kept glitching and messing up my team information," and a coder would put a '1' in 'negative' and a 1 in reliability and collaboration, and so on.

- Example positive valence: "I liked how fast the AR was."
- Example negative valence: "I'm worried about data security."
- Example neutral valence: "You'd have to consider server set up to roll this out in the field."

Table 7 outlines the absolute number of times an identity was indicated, as well as how often that identity had a positive, negative, or neutral valence. 124 comments were flagged as overall positive, 108 as negative, and 127 as neutral. The most leveraged identity was "Usability" 277 times (100 positive, 89 negative, and 81 neutral), followed by Logistics, Interaction, Efficiency, and Data Accessibility. The least leveraged identities were Privacy, Expense, Learning/Training, Other, and Information.

Identities with a larger 'positive' proportion were Communication, Usability, Immersivity, Speed, Collaboration, Interaction, and Learning/Training. Identities with a more 'negative' proportion were Trust, Flexibility, Adaptability, Reliability, and Information. Identities with a larger 'neutral' proportion were Logistics, Privacy, Security, Server, and Other. Identities that had no standout valence direction were Data Accessibility, Expense, and Efficiency.

Table 7 Semi-Structured Interview Coding

	Total	Total %	Positive	Pos%	Negative	Neg%	Neutral	Neu%
Total comments	359		124	**34.54%**	108	30.08%	87	24.23%
Positive	124		-	-	0	0.00%	1	0.81%
Negative	108		0	0.00%	-	-	0	0.00%
Neutral	127		1	0.79%	108	85.04%	-	-
Logistics	233	64.90%	70	30.04%	72	30.90%	87	**37.34%**
Privacy	9	2.51%	1	11.11%	2	22.22%	6	**66.67%**
Security	18	5.01%	3	16.67%	2	11.11%	13	**72.22%**
Server	61	16.99%	14	22.95%	16	26.23%	29	**47.54%**
Data Accessibility	102	28.41%	35	34.31%	31	30.39%	33	32.35%
Expense	15	4.18%	5	33.33%	4	26.67%	3	20.00%
Trust	34	9.47%	9	26.47%	14	**41.18%**	11	32.35%
Communication	98	27.30%	45	**45.92%**	28	28.57%	24	24.49%
Usability	277	77.16%	100	**36.10%**	89	32.13%	81	29.24%
Immersive	87	24.23%	35	**40.23%**	22	25.29%	28	32.18%
Speed	66	18.38%	27	**40.91%**	18	27.27%	21	31.82%
Efficiency	137	38.16%	48	35.04%	43	31.39%	43	31.39%
Collaboration	91	25.35%	45	**49.45%**	19	20.88%	26	28.57%
Flexibility	61	16.99%	12	19.67%	28	**45.90%**	19	31.15%
Adaptability	52	14.48%	5	9.62%	19	**36.54%**	17	32.69%
Reliability	61	16.99%	9	14.75%	47	**77.05%**	22	36.07%
Interaction	149	41.50%	59	**39.60%**	51	34.23%	36	24.16%
Learning/Training	12	3.34%	5	**41.67%**	2	16.67%	4	33.33%
Information	19	5.29%	3	15.79%	9	**47.37%**	7	36.84%
Other	15	5.01%	1	5.56%	6	33.33%	11	**61.11%**

We completed an exploratory factor analysis to see if there were emergent clusters that could further help describe the data. We selected the number of factors for clustering (eight) based on the number of factors with an eigenvalue above 1 (see Figure 26). We then set our factor loadings to suppress absolute loading values less than .3 to ensure that each identity within a factor would have a strong relationship within that factor. Our final result accounts 39.825% of the cumulative variance (well within our goal of 30-50%). Our significance tests revealed that are groupings had no common factors (chi sq. $p < .0001$) and that eight factors were sufficient (chi sq. 54.773, $p > .05$). Our factors were as follows:

- Factor 1: Privacy (.989) and Security (.379) [this factor explains 6.608% of variance]
- Factor 2: Collaboration (.741) and Communication (.689) [5.974%]
- Factor 3: Efficiency (.896) and Usability (.309) [5.501%]
- Factor 4: Speed (.976) [5.221%]
- Factor 5: Interaction (.733) [4.934%]
- Factor 6: Adaptability (.788) [4.405%]
- Factor 7: Data Access (.659) and Server (.362) [4.020%]
- Factor 8: Reliability (.535) and Trust (.444) [3.152%]
- Identities that did not strongly load on any factors were Learning/Training, Flexibility, Immersivity, Expense, Other, Logistics, and Information.

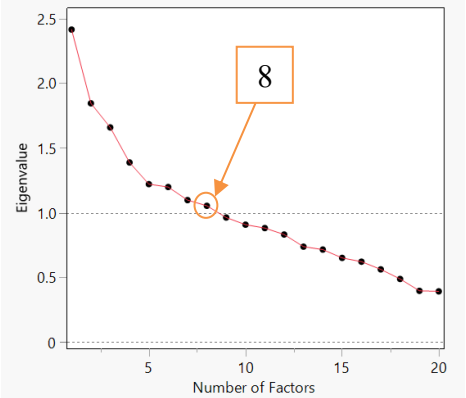

Figure 26: Scree Plot of Eigenvalues

To validate that we had the correct number of factors for our model, we evaluated a seven factor and nine factor model. The seven-factor model gave us more identities that did not load on any factors. The nine-factor model created inconsistent loadings with several identities spreading across two factors for only an additional 3% more of the variance being explained by that model.

4.5 Discussion

4.5.1 Qualitative Usability and Implications of the VTT from Responder Perspectives

To understand the role of the VTT during the simulations, we must explore the qualitative feedback provided by emergency responders. Responders had the most comments about usability, logistics, interaction, efficiency, and data accessibility. Responders were fairly evenly split with positive, negative, and neutral comments about the VTT and simulation experience but we provide a deeper, valence-based evaluation below.

Overall Strengths: Responders were most positive about the VTT regarding Communication (~46% positive), Immersivity (~40%), Speed (~41%), Collaboration (~49%), Interaction (~40%), Learning/Training (~42%), and Usability (~36%). Having the largest proportion of the valence (in relationship to the tri-direction of pos/neu/neg) indicates that while there is room for improvement, responders were more positive than not. This is especially true for Immersivity and Learning/Training, where their 2nd highest valence was neutral. We will define each of these terms based on the trends in these clustered identities.

- Collaboration refers to the feeling of teamwork being facilitated, while communication refers to responders being able to communicate with one another both synchronously and asynchronously. Responders felt that the VTT facilitated teamwork as they could look over the field and see the triage 'cards' of each patient, even when those patients were 40+ feet away. This, coupled with the ability to see the notes left by other responders on the VTT *"help[ed] with the ease of communication,"* which would be *"extremely helpful… for someone who is new to the incident."* The primary concerns (negatively valanced comments) centered on spotty wireless signal causing delay in system updates.
- Immersivity ended up split between two identities: Exploring whether the VTT helps/hinders immersion and discussing the simulations themselves. Responders indicated that the annotation features, intuitive use case, and having multiple responders coordinating facilitated the immersion. Regarding the simulations, responders liked *"to be able to hear their [the virtual patient's] breath"* and that the virtual patients had visible injuries that cannot be simulated on real people. The concerns surrounding immersivity had to do with glitchiness of the AR itself, like lagging hand tracking. Some responders indicated that they had a hard time knowing if the app had lagged or if their interaction had not successfully registered. The simulation took place in 95°F and caused some of the HoloLenses to overheat, which caused lagging issues.
- Speed also contained a split referring to either the VTT's ability to help triage or how quickly responders were able to pick up/learn how to use the VTT. Regarding the VTT's ability to assist in an event: *"I could see it [the VTT] cutting down on time for EMT's and paramedics," "I was able to triage a lot quicker… everything was streamlined,"* and "*AR would be extremely helpful in information handoffs".* Several responders liked the patient filtering feature as it enhanced how quickly they could determine the overall field status. Regarding the learning curve of the VTT, responders felt the AR was *"Easy and quick to learn*. Concerns with speed centered on occasional HoloLens lagging or the virtual elements jittering, forcing responders to close and re-open the tag and causing concerns about lost time.
- Interaction outlines what responders thought of the poking, raycasting, and voice interaction techniques available with the VTT, as well as the features within the VTT itself. One responder noted that they enjoyed the range of interaction options because *"… It was all embodied cognition*." Responders noted that *"Poke interactions were good for more accuracy and quicker than raycasting,"* and that they appreciated the annotation feature (voice). Concerns with interaction centered on whether the voice commands and annotation feature will work in loud ambient environments and that the redundant strategy, typing *"… in the AR [i]s clunky."*
- Learning/Training denotes how the responders feel about the VTT or the simulations for training purposes. Responders indicated that the 1V simulation offers *"real training advantages… It was a lot better to look and see what was on the patient. It's so hard to translate that on real patients when you've only*

practiced on people with no injuries on them." Moreover, *"AR is extremely useful for training. Using the AR for moulage (injury makeup) would help learning."* The 2R simulation and VTT together was considered *"... Highly useful for training, especially for mass casualty simulations"* because the 2R simulation affords recognition over recall and cognitive support, while the 2R simulation offers an environment to practice.

- Finally, usability describes how usable (or not) the VTT was. Responders were slightly more positive than any other valence on the usability experience of the VTT, though there were concerns. As usability touches many identities, we will talk more in depth about usability findings in its own section (4.5.4)

Broad Concerns: Responders were the most concerned (negative valenced) about Trust (~41%), Flexibility (~46%), Adaptability (~37%), and Information (~47.37%). These identities having a mostly negative valence indicates some deep concerns from responders, especially as the 2nd highest valence for all of these identities is 'neutral.'

- Trust straightforwardly denotes trusting the AR and VTT to work. We were unsurprised to see trust as with a high negative valence as responders self-describe as having low trust in new technology, stating *"I'm not huge on technology on the field… I want bare bone technology, that's enough for me," "If I have a sudden failure for whatever reason, a rain storm, whatever, all these tags I did are gone so now I have to re-triage."* Most of the trust concerns centered on HoloLenses failing during an MCI event, meaning some responders would not be able to see any of the triage statuses and would not have the benefit of a "*redundant system.*"
- Flexibility concerns the range of use cases the VTT has (between scenes), while adaptability outlines how the VTT can change within a given scene. Responders felt that AR and the VTT would only be meaningful for mass casualty events which do not happen very often. Moreover, as the MCI evolves, responders worry that any created applications will not be able to handle disturbance like events (see Chapter 3) and that it would *"... overbear the system if you have too many applications loaded onto it"* to compensate for the dynamic nature of MCIs. This finding is interesting as the VTT is made with our ecological analysis from Chapter 3 at its core. We designed the VTT to be flexible enough to work for any type of MCI scenario as long as there is sufficient wireless range for the server, battery power, ambient temperature, and brightness. While we acknowledge that those external environmental constraints are many, they are not to do with the MCI event itself. We posit perhaps these concerns are related to a low general trust in technology or an insufficient understanding of the technology (as too many applications will not overwhelm the AR system).
- Information discusses other information responders want access to or are concerned will be lost through using the technology. E.g., "*If an agency can't afford this device but they work with an agency who can,*

then this will create problems," and "*it is hard to track accountability in the AR… unless this mass casualty happens to be divided up along agency lines somehow."* This gets back at responders being concerned about whether the AR technology can successfully integrate on the field, indicating that there are fears that half integration is worse than no integration.

General Feedback: Responders also had mostly neutral thoughts about Logistics (~37%), Privacy (~67%), Security (~72%), Servers (~48%), and 'Other' (~61%). Generally, these were comments made by responders to let us know about any other concerns they would have with the VTT in the field. Each of these identities largely focused on other facets of MCI response than our simulations touched on.

- Logistics refers to actual implementation of the VTT in the field, concerns with implementation, and structures that needed to be considered. While logistics largely subsumes privacy and servers, there were times logistics touched on something unique like considerations about wayfinding, geotagging, 5G capabilities, radio integration, electricity, durability of the HoloLens's, etc. *"I would worry about after an explosion… having difficulties rendering if there's dust particles or smoke."*
- Privacy denotes maintaining patient privacy, while security is concerned specifically with bad actors accessing data they should not through the temporary servers.
- Servers were a component of the logistics needed to deploy the VTT in the field with responders questioning how that server set up would take place.
- 'Other' captured a handful of other ideas but was mostly concerned with how realistic the simulation patients were, saying *"patients could have been more interactive," "patients could have been more realistic,"* though responders did indicate that they would have preferred to do the virtual simulation earlier in the training day feeling like they "*could get [trained on MCI triage] faster.*"

Undecided: Finally, responders had no standout valence direction for Data Accessibility, Expense, and Efficiency, meaning they felt equally positive, negative, and neutral about those concepts in relationship to the VTT.

- Data accessibility describes how responders themselves will access data or share it with other teams/hospitals. These concerns were logistical in nature. Some responders felt positively that the VTT could facilitate cross-team communication, while other responders were concerned about how such a feat would be accomplished as each team would need the same headset, the same app, and to be on the same server, which all poses a different type of coordination issue.
- Expense is concerned with the cost of the AR devices (HoloLens 2's) in rural settings where response teams have limited to no access to funding, though responders felt that "Incorporating training into the AR [would] bring up the value," and that a training application is itself extremely valuable. They found

the synchronous communication between responders to be valuable and indicated that firefighter-based emergency response tools (like thermal imaging) would improve the value of the AR, stating that "*The more functionality it has, the more valuable it becomes."*

- Finally, efficiency gets at whether responders think the VTT helps with overall scene processing in an conformal way (as opposed to only evaluating speed). Responders liked that the VTT cut down on "*a lot of radio communication,"* that the VTT could "*help keep things organized"* and sped up determining overall field status. However, there were concerns about whether the VTT is actually quicker per-patient than "… just rip[ping] a paper triage tag off." This gets back to Chapter 2, where we stated that the real value of the VTT would be in the broader cognitive offloads and team tasks. In isolation or a small scenario, the efficiency of the VTT is dubious compared to a paper tag with a larger proportion of responders relative to patients in the field.

4.5.2 What Responders Care About

Unsurprisingly, some identities seemed to have association with other identities (like how data accessibility mentioned servers). Understanding how these associations cluster can help us understand what responders consider when confronted with a new technology. We did a factor analysis to determine if there were any overarching relationships and to that end, we found eight factors (or groupings) that helped explain the variability in the data. These factors underscore the primary considerations for responders and highlight the aspects that designers should focus on when developing new technologies for this context. By addressing these factors, designers can create tools that better meet the needs of responders, thereby improving overall effectiveness and efficiency in emergency situations.

1. **Factor 1 = Privacy and Security [Patient Data Security].** Responders consider: "Is sensitive patient data secure and safe?"
    - This factor emphasizes the importance of safeguarding sensitive patient information. Responders need assurance that any technology they use will protect patient data from breaches and unauthorized access. This is crucial not only for compliance with legal standards but also for maintaining trust in the technology.
2. **Factor 2 = Collaboration and Communication [Info-Based Synergy].** Responders consider: Am I able to access information from my peers and share information with my peers so we can work together?
    - Effective collaboration and communication are vital in emergency response scenarios. This factor highlights the need for technology that facilitates seamless information exchange among team members, ensuring that everyone has access to the necessary data to make informed decisions

and coordinate efforts efficiently. Information sharing needs to be quick and reliable with no lag or information drops.

3. **Factor 3 = Efficiency and Usability [Scale-Based System Efficacy].** Responders consider: How well does the tool itself work and how well does it work for the intended overall context?
    - This factor addresses the overall functionality and usability of the technology. Responders require tools that are not only effective in their immediate tasks but also scalable to handle the complexities of large-scale incidents. This includes intuitive interfaces and reliable performance under varying conditions. Thus, it is critical to evaluate how well a new technology or tool functions both at the tool level (See Chapter 2) and in the overall intended context (this study).
4. **Factor 4 = Speed.** Responders consider: Will this tool help me do my job quickly or will it slow me down?
    - In emergency situations, time is of the essence. This factor underscores the need for technology that enables responders to perform their duties quickly and efficiently, without unnecessary delays or complications (like from poor usability).
5. **Factor 5 = Interaction.** Responders consider: How does the new technology work and mediate my interactions with the field?
    - The way responders interact with the technology can significantly impact its effectiveness. This factor examines the user interface and experience, ensuring that the technology supports smooth and intuitive interactions that enhance, rather than hinder, the responder's capabilities. Developers must evaluate and understand how their technology works and mediates interactions in the intended use-case, evaluating both the deliberately designed interactions and unintended consequences of the technology
6. **Factor 6 = Adaptability.** Responders consider: Is the new technology able to adapt to my needs on the fly, as the MCI demands?
    - Emergency scenarios are dynamic and unpredictable. This factor highlights the importance of technology that can adapt to changing conditions and requirements, providing responders with the flexibility to respond to a wide range of situations effectively (acknowledge that MCIs are complex sociotechnical systems and provide agility).
7. **Factor 7 = Data Access and Server [Data Scaffolding].** Responders consider: If the technology acts as a center node for data transmission, what form does that take and how does it change my access to that data?
    - As technology increasingly serves as a central hub for data transmission and access, this factor examines the structure and organization of data within the system. Do not take the medium of

data transmission and access for granted. If the technology demands such a medium, that data hub will also have implications for design.

8. **Factory 8 = Reliability and Trust [Dependability].** Responders ask: Will this technology work 100% of the time and if not, will that be worse than never using the technology at all?
    - Reliability is critical in emergency response. This factor emphasizes the need for technology that functions consistently and predictably, minimizing the risk of failures that could compromise the response effort. Responders need to trust that the technology will work when they need it most. Designers need to ensure the hardware and software are both durable and reliable. Assess risks of potential failures and mitigate those risks or plan redundancies around failure. Without trust, responders will not adopt a new technology.

Identities that did not strongly load on any factors were Learning/Training, Flexibility, Immersivity, Expense, Logistics, Information, and Other. Some identities, such as Learning/Training, Flexibility, Immersivity, Expense, Logistics, Information, and Other, did not strongly load onto any specific factors. This can be attributed to their correlation with multiple identities, making it difficult for them to be categorized under a single factor. For instance, Flexibility, Expense, and Logistics were related to identities like adaptability, data access, collaboration, and server, making their unique categorization challenging. We hypothesize that Learning/Training and Immersivity did not load on any specific factor due to their primary focus on simulations rather than the Virtual Triage Tool (VTT) itself, or their broader application in training contexts. Consequently, we propose that Learning/Training and Immersivity could be considered as their own 'Simulation' factor, despite the lack of statistical backing for this claim. In the context of the valences, we see that responders were the most eager about the VTT's offerings within Factor 1 – Info-Based Synergy, Factor 4 – Speed, Factor 5 – Interaction, and 'Factor 9' – Learning/Training. There are mixed results on for the VTT on Factor 3 – Scale-Based System Efficiency (positive on Usability, undecided on efficiency). Responders were most concerned about Factor 8 – Dependability and Factor 6 – Adaptability. Responders had general feedback for 7 – Data Scaffolding.

4.5.3 Synergizing the Qualitative and Quantitative Data

To ascertain the effectiveness of the simulations and get a sense of the VTT's usability, we completed several within and between condition evaluations of p-score, error direction, time, and button presses. There were no significant within-subject findings, indicating that performance (time, error direction, p-score, button presses) in 1V did not provide any predictive power for performance in 2R. The biggest predictor of p-score, error direction, time, button presses, and overall accuracy was which condition participants were in. Unfortunately, the 1V (virtual patient) condition had poor accuracy at ~12% and participants were much slower than the 2R (real patient actor) condition that featured faster times and better accuracy at ~61%.

Even when correct, responders took significantly longer (~40 seconds) in the 1V condition to arrive at that conclusion than in 2R. The 2R condition had an error direction range of -1 to 3 with a mean of 0.07, while the 1V condition had a range of -4 to +4 with a mean of -0.64. This indicates that perhaps virtual patients were unclear representatives of their intended triage category. Responders were overall significantly slower in the 1V condition than in the 2R condition. In the 1V condition, correct and incorrect response times were not significantly different, but they were in 2R. Moreover, responders pressed significantly more buttons in the 2R condition than the 1V condition. All taken together, the time, accuracy, and button findings suggest that responders spent more time trying to interact with virtual patients to get necessary data and compensated for this by spending less time deliberating on the triage category or interacting with the triage tag.

When participants were correct in the 1V condition, they were ~8 seconds slower (though not statistically significant) than when they were incorrect. Comparatively, when participants were correct in the 2R condition, they were ~26 seconds faster (statistically significant) than they were when they were incorrect. Thus, correct responses took longer in the 1V condition while incorrect responses took longer in the 2R condition. We received feedback from responders that the virtual patients could be difficult to interact with due to their animations and could have been more realistic. Virtual patients would rock, hop on one foot, shake, etc. to indicate degree and location of injury. Unfortunately, this made it difficult for responders to get data readouts by putting their fingers on the moving virtual patient's wrist for heart rate, hand on the patient's bicep for blood pressure, etc. We posit this is why a correct answer took longer in the 1V condition. Comparatively, information was easier to obtain in the 2R condition which may explain why an incorrect answer took more time. E.g., when responders were unsure, they spent more time with a patient before ultimately making a choice to move on. The lack of relationship between error direction and time or button presses indicates that time spent with a patient does not predict a specific nature of error. When they were incorrect, responders tended to round toward the red (immediate) category. This means that when responders are unsure, they will play it 'safe' and put the patient in the highest priority category.

The 1V condition gave participant responders 10 minutes to triage 5 patients, while the 2R condition gave responders 10 minutes to tirage 20 patients in a team task. This team task dynamic yielded a standard emergent strategy for each responder to be responsible for 10 patients. No responders re-did their colleagues work during the simulation. E.g., there was no errant overlap as responders could see which patients had already been triaged. On average, 3.3 patients were triaged in the 1V condition, while 7.23 were triaged in 2R. Overall only 33 virtual patients were triaged across all participants, while 94 patient actors were triaged, representing a 184.84% increase. This increase in triaged patients is over the 100% increase in total patients for each responder to triage (from 5 to 10). We acknowledge that there is a limitation with our study design here. As this study was part of a bigger project to ascertain progressive training outcomes, these conditions

are not counter balanced allowing for there to be a practice effect. Moreover, there is an opportunity for a 'team' pressure effect where participants are motivated to perform better because they have a teammate who can see their performance synchronously. However, the accuracy of 1V being so much lower than 2R and the lack of a predictive relationship from 1V to 2R leads us to believe the difference between these conditions is less an issue with the lack of counter balancing and more a problem (or series of problems) with the 1V simulation itself.

Our findings reveal that the 1V condition may not be helpful as a practice simulation for responders than 2R, even though responders reported that they believe the 1V simulation would help them train faster. The 'moulage' realism of virtual patients was outlined as helpful for practicing effective triaging. However, responders also noted that the virtual patients could be more realistic which could help clarify the virtual patient's triage category. To help improve the time responders spent trying to gather virtual patient data, virtual patients need to be more interactive beyond special animations and unique interactions (such as listening to respiration rate). Virtual patients require either a larger hitbox for these interactions or a feature to slow down or pause their animation when a user is close to touching one of these hitboxes. Unsurprisingly, interaction time and button presses were significantly related but this relationship was not significantly changed by condition (condition 1V was about 1 second slower per button push). Any iteration of the 1V simulation will need to focus on improving accuracy and interaction time.

2R's accuracy rating >60% indicates that a 'real' patient actor simulation may be a viable training strategy for responders, though "*AR moulage"* or real moulage could further enhance the simulation. It appears that perhaps a 2R simulation is a critical step in conceptualizing the VTT as *"I was skeptical until we got to the actual live patients… and I was like okay, this is actually pretty useful."* Overall, when talking about AR for training, the responders focused on the 1V condition. Participant responders seemed disinclined to conceptualize the 2R simulation as a progressively context-sensitive simulation, instead focusing on the usefulness of the AR as a training, visual-cue guide for novice responders, stating *"This is much faster to learn than the VR system,"* indicating that responders considered the 1V simulation to be 'virtual reality' and the 2R condition to be 'reality.' This speaks to the immersion of the 2R simulation and also the profoundly reported use-case of AR for training.

While MCI response is too sensitive to test the VTT in the actual field, our study revealed that the virtual simulation (a more standard simulation strategy for academic evaluation) was not well suited to help us understand the actual efficacy of our VTT. This is, of course, due to limitations of our own development. However, the simulations were developed in partnership with a national training facility specializing in MCI triage. This underscores the complexity of creating meaningful simulations and highlights the criticality of 'real' field measurement (or as close as we can get at this stage of development). Despite the virtual

simulation's poor performance, however, a virtual simulation is still cheaper and (if made well and thoroughly tested) easier for both researchers and EMS stations to deploy for training purposes than the 'real' patient actor simulation.

4.5.4 A VTT Usability Deep Dive

To discuss the effectiveness of the VTT, we will focus on the 2R condition, as the 1V condition was affected by additional aforementioned usability issues that obscured overall performance measures. As noted, Usability had a slightly higher positivity valence (pos = 36.1%, neg = 32.13%, neu = 29.24%) indicating that the VTT had some strengths but also room for improvement.

In 2R, participant responders spent ~1 minute with each patient, indicating responders were able to use the VTT and still quickly move between patients. Responders took significantly longer with incorrect answers but button presses were not associated with accuracy. Together, this indicates that responders were spending more time with patients they were unsure about actually engaged in the cognitive triage task, not fighting with the VTT. Responders stated:

- *"I really like it. I think it's useful,"*
- *"I liked how well [the VTT] organized everything,"*
- *"[The VTT] would alleviate having to worry about losing your tag or having to write with a bloody hand,"*
- *"I could see it [VTT] cutting down on time for MCIs… so we're not having to stop and communicate 'did you check this person?"*
- *"As someone with an EMT background, I think it was really intuitive to understand,"* and
- *"I think knowing who's been treated and the time between [when they were last seen]… is a strong advantage exclusive to this system."*

Interactions (raycast, poke, and voice) afforded a redundancy in interaction styles. This meant responders could choose a preferred interaction technique, or pivot to a different technique if one was not working as hoped. Responders largely seemed to appreciate the interactions: *"Poking was the best way to do it… but for things further away, raycasting was really intuitive," "I used raycast for everything and I really liked it," "Both poking and raycasting were really intuitive, "Keeping your hands free (with voice) is a big benefit."* But there were responders who had difficulty using some of the interactions. *".. The buttons just weren't working for me," "Raycast was slowing me down," "more voice commands would be better,"* and concerns about a loud MCI event making voice commands unusable. It is possible that due to the heat, some responders had worse times with lag than other responders which gets back to the durability/reliability concerns outlined by responders. There were some issues interacting with the tag and its windows.

Responders indicating that the numpad for entering vital data sometimes would float into the middle of the triage tag, causing it to be essentially un-interactable. They would have to *"turn [their] head 90 degrees away and then back toward the tag to reset it."* Responders reported that "*I would like to be able to spam tap,*" meaning they want the programmed time between allowable button presses to be faster. This means there is a chance the VTT artificially slowed some responders down and that with some tweaking, response times might be even faster than ~1 minute per patient. It was reported that "It is a little disorienting sometimes to know if my application is running or not," due to windows loading out of their field of view or lagging inputs. Responders tended to round toward the red (or immediate) triage category when they were unsure. Future iteration of the VTT or other triage-supporting apps should help responders better discern between yellow, red, and grey.

Responders had some recommendations for improving the VTT experience, like a cueing system for when a window goes out of their field of view. More, they recommended some sort of loading indication or feedback for their interactions with the tag. The VTT does already provide some visual feedback (the pressed button becomes highlighted) but audio feedback could be added to provide redundancy. Responders indicated that the hand tracking could be improved upon, and that the HoloLens heating issues would need to be fixed before the VTT could be used in the field.

Overall, we have gleaned that the VTT has some substantial benefits and the drawbacks are mostly related to the AR technology itself, e.g., the battery life, durability of the device, tracking issues, overheating, and lag. For the tag (or any other MCI or emergency response AR tool) to reach real implementation, continued evaluation and hardware development will be crucial. We acknowledge that going into this study, we knew the VTT on HoloLens 2 would not be commercially viable yet. Responders agreed, and reported they could see AR deployed in the field *"Maybe in 5 or 10 years… when the kinks are worked out."* AR is an emerging technology that will continue to develop, and it is prudent to study its applications to help guide future AR hardware development and to give emergency response a running start once viable technology becomes available for the field. Any AR tool that reaches the field will need a heavily studied and validated background of research in order to assure reliability and safety, which could take years. Our current VTT is heavily informed, evaluated by, and iterated with emergency responders. Responders are eager about some of the unique strengths of the VTT and, with further iteration as technology develops, the VTT (or other AR apps like it) will be better situated to help patients and responders on a faster timeline.

4.6 Conclusion

This work is an extension of our prior work ascertaining the base usability of the VTT in a context-agnostic (i.e., lab) setting (Chapter 2). If the VTT is usable in more context-based simulations, this speaks to the

future applicability of the VTT in the field. Furthermore, this study contributes to an expanding body of research aimed at enhancing the utility and acceptance of AR in EMS. Successful application of AR in managing MCIs, which require the highest severity and time sensitivity of response, suggests broader feasibility of AR adoption within EMS. Not all MCI tools will be relevant in standard EMS response, but there are some shared and even unique applications for standard EMS like wayfinding, in-the-field guidance, context enhancers (showing pipelines, etc.), and more. We found that the VTT offers some profound, unique benefits by facilitating communication, collaboration, and learning. We also found that there are still many hurdles to overcome before we see AR in the field. Responders were able to provide critical, expert feedback and gave recommendations for ways to improve the tag (allow for faster button clicks, fix the drift on the numpad, etc.)

This work also builds our understanding of the usefulness of our virtual simulation relative to the 'real' patient actor simulation, with implications for future MCI training (particularly for rural responders) and other general EMS training. There's potential for the EMS field to begin with AR-based training to stay current on best medical practices and, as AR matures, to extend AR into future field applications. One of the greatest hurdles with new technology in EMS is responder acceptance, which these simulations could also help with. However, we note that despite developing our 1V (Fully AR) simulation in partnership with subject matter experts in MCI triage training, we still encountered major problems in rendering a meaningful triage experience for responders, underscoring the complexity of simulation development and highlighting the need for more 'realistic' simulations (like our patient actor simulation). Our study provides rich data about MCIs and EMS as the participants are responders themselves.

4.6.1 Limitations

This work features some limitations. To start, we acknowledge that we only have thirteen participants. We argue that these findings are still meaningful because all of these participants are actual EMS practitioners. Moreover, the intensity of our 'real' simulation is, to our knowledge, a first of its kind for AR MCI evaluation. There is a possibility of order effect as the virtual simulation was always before the 'live' simulation. This project was part of a much larger study determining triage training efficacy, thus we could not counter balance the order of tasks. Due to the intense summer heat in Texas, we ran into overheating issues with the HoloLenses which caused some lagging issues, muddying some of our usability findings. Finally, we acknowledge that our measure of the VTT's effectiveness is confounded by whether responders have a solid handle on triage itself, as well as on AR interactions. We tried to minimize this effect by requiring each participant to be a responder and by furnishing extensive triage and AR training for several hours before simulation testing. Finally, ideally, we would have a 'control' condition with responders completing triage without the use of AR and instead with standard paper triage tags. This study was a large, 1-week

long event with large time requirements as it was, but we acknowledge that this means we cannot directly compare the success of the VTT to the current 'standard' during triage. Finally, we acknowledge that this work did not provide an interrater reliability evaluation. Reliability measurement would be ideal, and will be done in future work with more nuanced splits in a-priori identities. E.g., "speed" will be broken up as two separate identities for speed of learning the AR and speed of processing the MCI scene (as informed by the analysis in this work thus far).

4.6.2 Future Directions

Future work should gather more participants and counter balance the virtual and live simulations. Future work should also leverage the NASA TLX and SUS to measure qualitative perceptions of the VTT better quantitatively across contexts and the simulations themselves. The 1V simulation needs iteration and re-testing to evaluate its actual viability as a training strategy for emergency responders, especially given the profound responder interest in such an application. Real implementation of any AR tools in the MCI field will require training even if just to learn the AR itself, exacerbating the need for solid 'virtual' simulations. Future work should assess a 'real' simulation like 2R for triage accuracy via the standard paper tags in order to create a more direct comparison between the AR and current best practice. Finally, more goes into triage accuracy than understanding how to use the SALT framework, AR, and/or the VTT. Given that this AR context of MCI triage involves supporting human-to-human interactions, future work should also explore a discussion of triage performance (accuracy/time, p-score, error direction) based on social factors like bias.

5 Study 4: Exploring the Impacts of Bias on Mass Casualty Incident Triage: Trends across Virtual and Real Patients in Augmented Reality

5.1 Introduction

Mass casualty incidents (MCIs) overwhelm available resources with deadly consequences due to "… poor leadership, planning, communication, and resource management" [1]–[4]. Triage is a core MCI management strategy, a subsystem used to approach the scene medically, and requires intensive planning, communication, and resource management to ensure medical resources are applied with the most impact [96]. This 'impact' is determined based on injury severity and available resources, and must be done expediently and accurately to facilitate heightened treatment efficacy [7]. Worded differently, triage assesses and prioritizes patient injuries to determine who gets what resources, faster care, and priority evacuation [5], [6]. MCI triage imposes intense cognitive demand and time pressure in a dangerous environment, often with responders having limited to no prior experience in applying MCI triage skills, which is exacerbated by potential skill degradation between MCI events [81]. This cognitive demand comes from a heightened requirement for situation awareness entailing scene development and complex conformal relationships between responders, available resources, casualties, evacuation procedures, hazards, and scene clearance. Unfortunately, all these pressures have uniquely negative impacts on people of color [8]–[13]

There is extensive evidence of anti-black bias in healthcare [8]–[13] that manifests as slower, lower quality, and less accurate care yielding significantly worse health comes for black patients (even when socioeconomic status is accounted for) [8], [9], [13], [17]–[19] as a result of both implicit biases and institutional racism. At the larger end of the scale, *Institutional racism* is the embedding of racism in policy and practices that give advantages to some racial groups while disadvantaging others [9], like hospital policies that disproportionately negatively impact black patients. *Cultural racism* can happen more personally, like between a healthcare provider and patient, and encompasses discriminatory behavior and decisions against black patients. This negative change in behavior toward black patients can be conscious/overt, or unconscious/unintentional [9], [109]. *Implicit biases* are the unconscious, unintentional implicit associations between a group and a negative evaluation [15]. The more cognitive load a person is experiencing, the more they rely on biases to provide 'snap judgements' to respond quickly. The adverse effects of implicit bias are correlated with the level of the emergency and, therefore, worsen with emergency event severity [20] due to increasing time pressure, patient volumes, and diminishing information [8], [21], [22], which each yield more demanding cognitive load. The compounding effects on cognitive load make emergency contexts "… breeding ground[s] where implicit biases particularly influence clinical decisions" [8].

Specifically, triage is slower and less accurate for black patients during MCIs [23], [24], which critically impacts health outcomes during the 'golden hour' of triage.

In Chapter 2, we designed an augmented reality (AR) virtual triage tag (VTT) to facilitate MCI triage that was then evaluated via representative tasks in a lab setting to determine the VTT's usability compared to the standard paper-triage tag [25]. We extended this work in Chapter 4 to evaluate effectiveness and emergency responder (ER) perceptions of the VTT in more context-sensitive virtual (1V) and 'real' patient actor (2R) simulations. We also evaluated the effectiveness of two MCI triage simulations. We found that the 1V condition had significantly depreciated accuracy relative to 2R, despite having no significant differences in time spent per patient or button presses. The future work of that chapter outlined a need to consider more social facets of performance, like racial bias, that could impact the time spent with each patient, accuracy, and types of errors made.

In applied settings, AR is often not the 'primary task,' meaning AR instead should support a real-world primary work task like construction [110], bridge inspection [111], driving [112], etc. and not distract. The AR is 'referred to' or used as a tool to facilitate the primary task (this is different from education [113] and rehabilitation [114], where the primary task is to be educated/rehabilitated entirely through AR interaction). Notice that these primary tasks do not involve direct interactions with other people. However, we must consider that a large subset of potential work tasks, including those that do or will leverage AR, involve working with people directly, like MCI triage. A core part of determining the effectiveness of any human-to-human AR support entails thoughtful implementation and evaluation of patient diversity within the simulation due to the aforementioned impacts of implicit racial bias in healthcare. Findings from such an evaluation has implications for future studies and simulations because 1) an ability to detect racial bias in a simulation elucidates the need for robustly diverse study stimuli, and 2) outlines opportunities to capture and mitigate bias impacts during their training simulations, long before the responder is in the field with real patients. As aforementioned, Chapter 4 evaluated a virtual (1V) and 'real' (2R) patient actor simulation. This chapter (chapter 5) explores bias as a potential predictive factor for accuracy because there is the potential for bias-based variation in emergency response triage. Thus, we want to e*xplore if racial bias manifests in the simulations and its implications for emergency response and future diversity in study design.* To this end, using the performance measures defined in Chapter 4, we will explore how racial bias manifests in the VTT and its implications for emergency response and future diversity in study design. More specifically, we want to understand:

8. What are the impacts of racial bias on performance across virtual (1V) and 'real' (2R) simulations?
9. Does racial bias impact the type of error classification?

Looking back at the second wave of the triage flow chart (Figure 17), the only difference between an immediate (red) and expectant (grey) category patient is the question: "Is the patient likely to survive given current resources?". This is a more subjective question than it might appear at first. Critically, this one question determines who gets priority care and who gets left on the field to expire. As patients of color have their pain and ailments taken less seriously [115] which yields worse health outcomes, we want to determine more than overall triage accuracy. We want to evaluate the types of errors being made by responders and in what directions. E.g., are black patients who should be immediate (red) more likely to be assigned grey and left behind, and are white patients who should be delayed (yellow) more likely to be assigned red and given priority consideration? If so, future training using these simulations could infuse mitigation strategies. Thus, we leverage the Implicit Association Task (IAT) to capture implicit bias and compare those scores with overall triage performance on two simulations with racially diverse patients.

If racial bias can be measured and determined before 'live' training, then virtual simulations have a new and compelling justification. Responders can learn updated best practices for the field, while also critically engaging with more pernicious facets of effective response in a safe environment for both them and the real patients potentially impacted by their biases downstream. Moreover, if virtual and 'live' simulations can capture racial bias impacts, then this study details the criticality of diverse stimuli for both simulations and for future studies in AR human-to-human supportive tasks. This work expands on a growing body of knowledge showing how diversity study stimuli can influence the way people perceive each other [116], [117] and participant performance [118]. Though we do not have IAT results for gender, we do know that intersectional identities like black women have even worse outcomes [115], [119], with white women having depreciated outcomes relative to white men [120]. Thus, this work also evaluates the impacts of gender, and gender*race on participant performance. We hypothesize the following:

- There will be differences in triage performance (p-score, time, accuracy, and button presses) based on patient race
- These differences will be related to triage severity. E.g., when white patients are triaged incorrectly, tend toward a 'more severe' category. When black patients are triaged incorrectly, they will tend toward a 'less severe' category.
- White patients will get more time allocated to them than black patients
- There will be differences in triage performance (p-score, time, accuracy, and button presses) based on patient gender
- These differences will be related to triage severity. E.g., when male patients are triaged incorrectly, they will tend toward a 'more severe' category. When female patients are triaged incorrectly, they will tend toward a 'less severe' category.

- Male patients will get more time allocated to them than female patients
- There will be differences in triage performance (p-score, time, accuracy, and button presses) based on patient gender*race
- These differences will be related to triage severity. E.g., when incorrect, white male patients will be triaged 'more severely' than black female patients.

This work also considers the impact of participant motivation on their triage performance.

- We hypothesize that more motivated participants will be more accurate
- More motivated patients will be faster

To our knowledge, our work is the only example of measured racial bias compared against responder performance in MCI triage, thus providing a necessary and rich data set. Moreover, this is the only example measuring racial bias 'in the field' as opposed to in the emergency room after patients have been evacuated [23], [24]. We describe the biases we found in the data, as well as some concepts that may help explain our findings.

5.2 Methodology

5.2.1 New Measures

This chapter evaluates the same time to completion and accuracy measures as in Chapter 4. This time, however, we additionally explore error type, motivation, participant age, and implicit association task (IAT) scores.

- At the beginning of the study, participants were asked to rate how motivated they were to participate on a scale of 1-10 with 10 being the most motivated. We captured motivation and participant age as the IAT can be affected by other factors and we wanted to explore if motivation or age may be correlated with it.
- The IAT was used to measure racial bias at the beginning of the study.
- Error type is a qualitative analysis of the exact types of errors captured in error direction.

These are the independent and dependent variables for this study:

IV	**DV**
Motivation	Time
Participant Age	P-score
IAT	Error Direction
Triage Category	Button presses
Patient Gender	Overall Accuracy
Patient Race	
Patient Gender*Race	

The IAT was invented by Greenwald et al., in 1998 [121], [122] with Greenwald continuing to refine the task (now only requiring five trials instead of seven) and statistical evaluations of task performance since [123]–[126]. There are limitations with the IAT, namely, that the IAT may be influenced by factors unrelated to actual implicit biases [127] and there are questions about whether the IAT can actually predict real-world discriminatory behavior [128]. However, in social psychology, the IAT is considered an indirect, unobtrusive measure critical for capturing context-sensitive attitudes [129] that is still receiving regular updates from its original creators [124]. Our context is MCI triage and the biases among emergency responders which is a sensitive context, thus we strive for a context sensitive measure. The IAT has been used hundreds of times [122] to measure implicit biases, including racial bias. This is done by measuring the strength of associations between groups. For racial bias, this is done with four stimuli types: black/white faces and positive/negative words. We leveraged Dr. Meade's IAT open-access software [130] to furnish the IAT (seen in Figure 27). Stage one (A) in Figure 27 entails learning to associate the images of white or black faces with the appropriate side of the screen labeled "African American" and "European American" via the 'e' key for left and 'i' key for right. (A) depicts an 'incorrect' response. Stage two (B) is where participants learn to encode the valence of the word stimulus with a side of the screen as well. Stage 3 (C) is the first paired test trial where participants are asked to assign the face or word to the appropriate side of the screen. Notice that "African American" is on the same side as "positive words." Participants then go through stage 4, a retraining where one of the categories changes (e.g., for our purposes, the side of the screen for African American vs. European American switches).

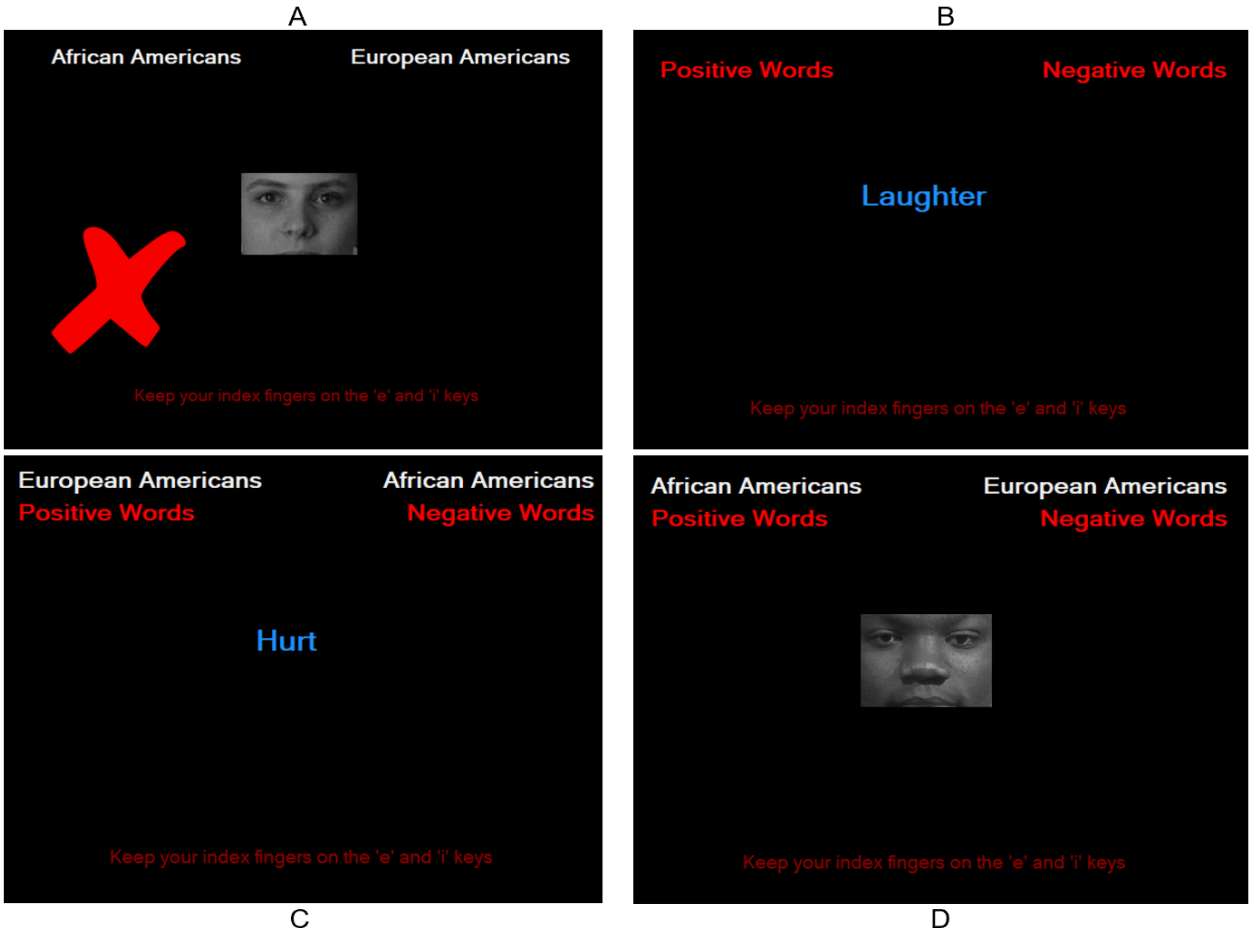

Figure 27: IAT stages 1, 2, 3, and 5 using Meade's IAT software [130].

Then, participants completed stage five, or the second paired test trial (D). Response time (in milliseconds), stimulus name, and accuracy are recorded for each trial. Each stage (or block) is 30 trials long, and this data informs the final 'IAT score' (calculation defined in the results). The response times and accuracy are used to create a GNB (bias) score.

5.2.2 Virtual vs. Real Simulations, Tasks & Procedure

The IAT was given at the beginning of the participants' day just after they completed the consent forms. This study used 'deception' by omitting that this study was capturing or measuring racial bias. This was done to ensure participants would not mediate their engagement with the IAT or with the racially diverse patients. Participants were debriefed at the end of the study and given the opportunity to withdraw their data (zero chose to do so). The virtual simulation featured 2 black men, 1 white man, 1 black woman, and 1 white woman. The 'live' simulation featured patient actors as follows: three men of color, six women of color, two white men,

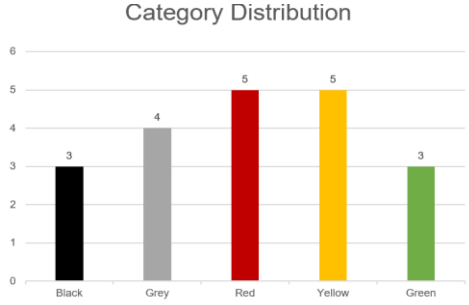

Figure 28: Triage category distribution

and eight white women. In chapter 4, we indicated that triage categories in the 'live' simulation were not equally represented because green and black would be easy to determine quickly and we cared more about the nuances between the other categories. This breakdown can be seen in Figure 28.

5.3 Results

At its core, this is the same dataset used in Chapter 4 (time, accuracy, p-score, button presses, and error direction) with some additional, yet un-analyzed data (motivation, age, IAT scores, error direction). This analysis examines the Chapter 4 data in new ways with the addition of these new data to explore impacts of IAT bias on triage performance. Moreover, this analysis now considers participant age and features some exploratory analysis on whether patient gender impacts performance. The same missing data considerations outlined in Chapter 4 are utilized here for these analyses. We first describe the data and some basic relationships in section 5.3.3, and then detail a mixed methods model for more complex interaction effects in section 5.3.4. These missing data and the complete-case consideration can impact degrees of freedom between tests.

5.3.1 Participants

This study utilizes the same participant pool as our previous work (Chapter 3) and includes 13 participants. These 13 participants comprise 2 female and 11 male participants (15% and 85% respectively) with an average age of 36.23 years (with a range of 21 to 60). All participants identified as White/Caucasian. Among them, ten had EMT-Basic or EMT-Paramedic level certifications, two had broad experience in emergency response, and one had CPR and First Aid Training. Additionally, ten participants had previous MCI training, six indicated three or fewer hours of AR/VR experience, and two indicated five or more hours of AR/VR experience. In terms of representativeness, approximately 85% of responders nationally are White/Caucasian, about 2% are Black/African American, and around 76% are male, with an average age of 38 years [105].

5.3.2 Calculating IAT

Meade's IAT software automatically generates an overall IAT score (GNB) through a process defined by Greenwald, Nosek, and Banaji in 2003 [125]. The process is as follows [131]:

- Trials with reaction times >10,000ms are eliminated and participants with 10% or more of their trials being <300ms are eliminated.
- Mean correct response time is calculated via the correct answer response times for blocks 3 and 5 separately.

- Then, response times for initially incorrect items are replaced with the block mean of correct items for their respective blocks +600ms.
- Corrected averaged response times for the 3 and 5 blocks are now calculated.
- Standard deviation (or pooled SD) is calculated as the standard deviation of all correct/incorrect responses across blocks 3 and 5.
- The final GNB score is the average corrected response time from block 5 minus the average corrected response time of block 3, with this difference being divided by the pooled SD.
- A negative score means items in block 5 have a STRONGER association than those in block three. For our IAT, a negative score means there is a preference for the African American → negative word / European American → positive word pairings [132].

From here, 'categories' are determined based on the final GNB score. A score >±.15 means there is a minor bias, >±.35 = moderate bias, and >±.65 = strong bias. The nuance comes in the positive or negative valence of the score. In our case, a negative score means a negative black (positive white) bias whereas a positive score means a positive black (negative white) bias. One participant had a negative black bias, seven had a moderate negative black bias, two had a minor negative black bias, two had no bias in either direction, and one had a minor positive black bias.

5.3.3 Data Characteristics

We checked the 'new' data we are using for secondary analysis (motivation, age, and IAT) for outliers with quantile plots and a Huber M-Estimation set to 4 to balance efficiency and robustness and found no outliers. Then we checked the distribution of these variables and found them all to have a skewness between -2 to +2 and kurtosis between -7 to +7, meaning these data are normally distributed. Of note: 'deceased' patients in the 'Real' simulation were mannequins and were excluded from our analysis. From here, we ran several univariate ANOVAs to explore some direct relationships between the variables:

- The total number of triaged patients was 33 in the 1V condition. ~42% of the triaged patients in 1V were a person of color (14). 94 patients in total were triaged in the 2R condition, of which ~35% were a person of color (29).
- There was a significant effect of IAT score on overall accuracy when we combine both conditions [$F(1, 78) = 8.9121$, $p = .0038$] and by condition 2R [$F(1, 107) = 12.5652$, $p = .0006$]. Overall accuracy refers to the overall accuracy rate by participant for their entire 1V or 2R session. On the spectrum from strong negative black bias to strong positive black bias, more negative IAT scores yielded lower overall performance.
- There was a significant effect of motivation to participate in the study on overall accuracy only for the 2R condition [$F(1, 67) = 47.9267$, $p < .0001$] with higher motivation yielding higher

performance. Though there was no significant effect of motivation on time for the 2R condition, there was for 1V [$F(1, 25) = 18.1798$, $p = .0003$] with lower motivation yielding slower times.

- In condition 2R, there was a significant effect of age on error direction, with older participants being more likely to round toward a more severe triage category $F(1, 82) = 5.2875$, $p > .05$. There was no significance for condition 1V.
- There was no significant effect of age on accuracy for either condition.
- In condition 2R, there was a significant effect of age on duration with older participants taking longer [$F(1, 82) = 6.6030$, $p < .05$]. There was no significant effect of age on duration in condition 1V. Similarly, there was a significant relationship between age and button presses for condition 2R, [$F(1, 82) = 5.4297$, $p < .05$] with older participants pressing more buttons; but not for condition 1V. However, age only had a practically meaningful effect on time (1.16 extra seconds per year), with button presses effect size being rather small (0.09 extra button presses per year).
- We did not find any significant effects of patient gender on time, button presses, accuracy, or error direction. However, we only had two female participants so those analyses were solely exploratory in nature.
- There was a significant effect of motivation on IAT [$F(1, 95)= 20.8425$, $p <.0001$]. More motivated participants had lower anti-black bias.
- There was a significant effect of age on IAT [$F(1, 107) = 26.4609$, $p<.0001$]. Older participants had lower anti-black bias.
- There was a significant effect of age on motivation [$F(1, 95) = 22.8554$, $p <.0001$]. Older participants were more motivated to participate in the study.
- An exploratory model examining the relationship between age, motivation, and age*motivation on IAT scores was done. The model had poor fit ($p<.0001$), but to assess multicollinearity, we evaluated the Variance Inflation Factor (VIF) for each parameter estimate and found each to be <5, indicating there is not a multicollinearity issue between these factors.
- An exploratory model examining the relationship between IAT, age, and motivation on overall accuracy was done. The model had good fit and was significant [$F(3, 93)=9.8742$, $p<.0001$] but only IAT was considered a significant predictor.
- There was a significant effect of motivation on button presses ($p<.05$) with more motivated patients pressing fewer buttons.

In section 5.2.2 we outlined the distribution of available triage categories. In the 1V condition, there is one patient for each triage category for condition. For condition 2R, there are 3-5 patients for each category in (see Figure 28). As the task was to triage as many patients as possible in the time constraint, the distribution of actually attended to patients was captured. Table 8 outlines the number of each category triaged across

all participants and the overall accuracy of participants for that category. The types of errors can be seen in Figure 29.

Table 8 Errors by Category

	1V	1V accuracy	2R	2R accuracy
Total Triaged	33	12.5%	94	60.76%
Deceased	3	0%	10	100%
Delayed	14	10%	26	60%
Expectant	0	n/a	23	61%
Immediate	4	50%	20	60.7%
Minor	10	10%	15	61.5%

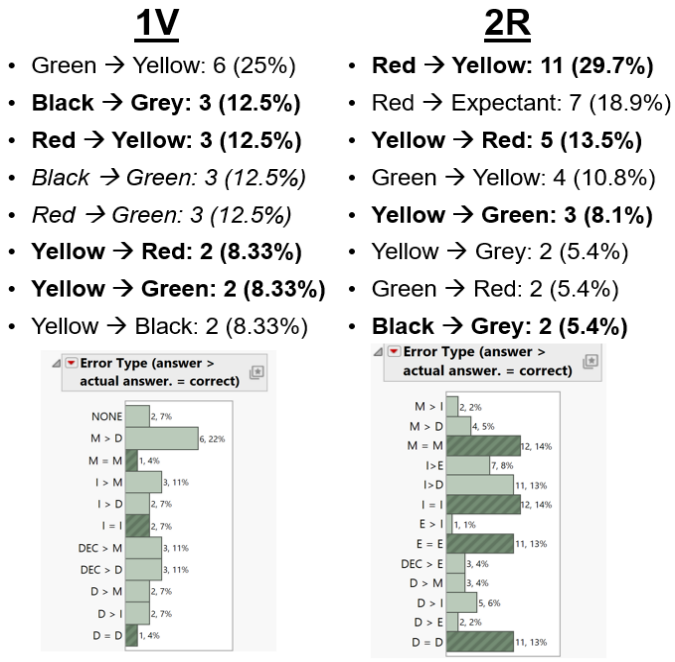

Figure 29: Histograms of Error Types by Condition

5.3.4 Constructed Models and Results

To determine impacts of racial bias on performance across virtual and real contexts, we ran four standard least squares regression models to explore the main and interaction effects (detailed in the bulleted list below) on p-score, error direction, time, accuracy, and button presses. Model 1 provides a direct effect analysis of each dependent variable by condition; model 2 includes interaction effects (established a priori) to explore whether these give us a better fit; model 3 explores an overall direct effect model including both conditions; and model 4, another overall model including both conditions but this time exploring if interaction effects give us a better model fit. Unfortunately, due to the lower degrees of freedom in the 1V condition, it was not possible to test interaction effects for model 2 condition 1V without significant biased and zeroed parameter estimates.

The interaction effects were established a-priori to evaluate the impacts of IAT bias on race-based triage performance, time, etc. These analyses afford a deeper dive into the category-based accuracies and error direction. This helps us evaluate whether white patients are given more priority than they should or black patients less priority, etc.

To explore what factors may explain these performance differences other than race, we also include exploratory analyses based on patient gender.

Main Fixed effects (all models):

- Race category of the patient (white, mixed, black)
- Actual triage category of the patient (minor, delayed, immediate, expectant, deceased)
- The participant's IAT score based on category (minor, moderate, strong)
- Patient sex (M/F)
- Participant Age
- Participant Motivation (1-10)

Fixed interaction effects model 2:

- Patient race*IAT category
- Patient race*Actual triage category
- Patient sex*Actual triage category
- Patient sex*Patient Race

Fixed interaction effects model 4:

- Patient race*IAT category
- Patient sex*Patient Race
- Patient Race*Patient Sex*IAT categories

Condition 1V, Model 1:

PERFORMANCE (P)-SCORE: There was no significant effect of IAT category ($p = .7709$), triage category ($p = .2927$), patient sex ($p = .5369$), patient race ($p = .1150$), participant age ($p = .1876$), or motivation ($p = .4829$) on p-score [see Table 9]. The model had an insignificant lack of fit test ($F = 158.2301$, $p = .0624$) indicating that the model fits the data adequately but the large F and trending p value mean the fit could be improved. The large Max RSq (.99), small adjusted R square of .05, and insignificant ANOVA further indicate that the model does not explain the variability on p-score and could be improved. To assess

multicollinearity, we evaluated the Variance Inflation Factor (VIF) for each parameter estimate. All VIFs were < 5 except for motivation (VIF = 9.052). While this 9 is below the standard 10 cutoff, we explored the data more. The only other variable approaching a moderate collinearity problem was IAT category (VIF = 4.107). A post hoc correlation matrix of motivation on IAT score revealed an Rsquare of .23 and significant ANOVA [F(1, 26) = 7.930, p = .0091]. Participants with higher racial bias against black faces had lower motivation. Removing motivation from the model lowered all VIF's below 3 but did not make the model significant [F(9,19) = 1.2292, p = .3] or drastically improve model fit (adjusted R square from .03 to .088).

Table 9 – Model 1 1V Performance P-Scores Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	0.9563	1.3488	0.2927
IAT Categories	0.0937	0.2642	0.7709
Patient Sex	0.0704	0.3971	0.5369
Participant Age	0.3341	1.8847	0.1876
Patient Race	0.489	2.7589	0.115
Motivation	0.9122	0.5146	0.4829

ERROR DIRECTION: Patient sex was the only significant predictor on error direction (p = .0120) with women having an average error direction of .7 and men -1.74 [see Table 10]. This means men are more likely to be given a more severe triage category while women are more likely to be given a less severe category than what they should have been given. This model had good fit (F = .1909, p = .96) and explained a significant amount of the variation [F(9, 15) = 6.2562, p = .001] with an adjusted Rsquare of .66342. VIF was again evaluated and again found all variables < 5 except for motivation (9). While 9 is below the standard threshold of 10 we explored removing motivation from the model and found that doing so made the model worse (adjusted Rsquare fell to .5937).

Table 10 - Model 1 1V Error Direction Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	1.4455	0.4374	0.7295
IAT Categories	5.4403	2.5098	0.1148
Patient Sex	**8.9838**	**8.1542**	**0.012**
Participant Age	3.0325	2.7525	0.1179
Patient Race	3.3428	3.0432	0.1015
Motivation	1.1219	1.0183	0.3289

RESPONSE TIME: Motivation was the only significant predictor of time spent with each patient (p = .0066). Goodness of fit test (F = 28, p = .14) indicated sufficient model fit, while the ANOVA [F(10, 16) = 2.88, p = .0289] indicates that the model predicts a significant amount of the variation in response time

(adjusted Rsquare of .4197). Older participants were slower. When we remove motivation from the model due to its 9 VIF score, this again worsens the model (adjusted Rsquare falls to .1245) and renders it non-significant. See Table 11 for full numbers.

Table 11 - Model 1 1V Response Time Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	6980.5	1.3263	0.3027
IAT Categories	7959.541	3.0246	0.0769
Patient Sex	2329.036	1.7701	0.202
Participant Age	64.059	0.0487	0.8282
Patient Race	1110.502	0.844	0.3719
Motivation	**12798.879**	**9.7272**	**0.0066**

BUTTON PRESSES: IAT categories (p = .0006) and motivation (p <.0001) were significant predictors of button presses [see Table 12]. Model fit was good (F =12.51, p = .2187) and significant [F(10, 16) = 4.8771, p = .0026] with an adjusted Rsquare of .59854. Participants with a strong negative black bias pressed fewer buttons than those with a moderate bias, and moderate bias still pressed fewer buttons than the no bias category. Interestingly, less motivated participants pressed more buttons. Again, removing motivation from the model made the model worse (adjusted Rsquare fell to .01).

Table 12 - Model 1 1V Button Presses Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	0.29942	0.0127	0.9996
IAT Categories	**142.0076**	**12.0392**	**0.0006**
Patient Sex	0.91359	0.1549	0.6991
Participant Age	0.93133	0.1579	0.6963
Patient Race	5.0198	0.8511	0.3699
Motivation	**160.0496**	**27.1375**	**<.0001**

Condition 2R, Model 1:

PERFORMANCE (P) SCORE: Motivation (p = .045) was the only significant predictor of p-score in the 2R condition [see Table 13]. The model had good fit (F = 1.4542, p = .257) but did not explain a significant amount of the variation in p-score (adjusted Rsquare of .0368). The max Rsquare of .886 indicates there's potential for the model to be improved. Unlike in the 1V condition, none of the parameters had a VIF above 4.

Table 13 - Model 1 2R Performance (P) Score Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	0.5835	0.7028	0.5543
IAT Categories	0.2341	0.2819	0.8382
Patient Sex	0.0263	0.0952	0.7588

Participant Age	0.2077	0.7502	0.39
Patient Race	0.2025	0.3658	0.6953
Motivation	**1.1633**	**4.2032**	**0.045**

ERROR DIRECTION: Triage category (the triage category the patient is representing) was the only significant predictor of error direction (p = .0286); see Table 14. Generally, responders tended toward the red (or immediate) triage category. E.g., the most common errors were to over assign yellow (delayed) patients as more severe or to move grey (expectant patients) to a less severe but higher priority category. However, despite a good model fit (F = .9716, p = .5628) the model did not predict a significant portion of the variation (11, 57; F=1.7965, p = .076) despite an adjusted Rsquare of .11. VIF scores were all < 4.

Table 14 - Model 1 2R Error Direction Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	**4.8931**	**3.2404**	**0.0286**
IAT Categories	0.9365	0.6201	0.6048
Patient Sex	1.0773	2.1402	0.149
Participant Age	0.219	0.435	0.5122
Patient Race	0.5482	0.5446	0.5831
Motivation	0.0513	0.1019	0.7508

RESPONSE TIME: Age was the only significant predictor of time (p = .0024, see Table 15) with a significant effect [ANOVA (11,57); p = .0004] and an adjusted Rsquare of .31. However, the goodness of fit test was significant (F=2.5866, p =.0453) suggesting this model does not fit the data well. VIF scores were all < 4.

Table 15 - Model 1 2R Response Time Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	735.524	0.1799	0.9096
IAT Categories	7254.857	1.7741	0.1624
Patient Sex	2425.543	1.7795	0.1875
Participant Age	**13789.541**	**10.1164**	**0.0024**
Patient Race	5750.784	2.1095	0.1307
Motivation	5023.766	3.6856	0.0599

BUTTON PRESSES: IAT category was the only significant predictor of button presses (p = .0082) [see Table 16]. Responders with a minor pro black bias pressed more buttons than any other bias category. Motivation was approaching significance (p = .0603) with higher motivation yielding fewer button presses. The model had good fit (F = 1.5241, p = .2289) and was significant [(11, 57), F = 3.4840, p = .0009), predicting 28.66% of the variability (adjusted Rsquare). VIF scores were all < 4.

Table 16 - Model 1 2R Button Presses Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	19.1924	0.4332	0.7301
IAT Categories	**191.5084**	**4.3224**	**0.0082**
Patient Sex	1.6484	0.1116	0.7395
Participant Age	18.6686	1.2641	0.2656
Patient Race	15.3931	0.5211	0.5966
Motivation	54.2379	3.6725	0.0603

Condition 2R, Model 2:

PERFORMANCE (P) SCORE: None of the main or interaction effects were significant predictors of p-score. The model had good fit (F = 1.4883, p = .2456) and a good Rsquare, but the adjusted Rsquare (.035) indicates the model is too complex and needs improvement. VIF scores were all < 5.

ERROR DIRECTION: None of the main or interaction effects were significant predictors of error direction. Again, the goodness of fit test was not violated (F=1.2464, p = .2582) but the model had a low adjusted Rsquare (.0769) and insignificant ANOVA (23, 45; F=1.24, p = .2582).

TIME: Age was the only significant predictor of time (p = .0018). However, this model violated the goodness of fit test (F=3.1285, p <.05) indicating that this model does not fit the data well.

BUTTON PRESSES: IAT categories was the only significant predictor of button presses (p =.0057) with pro black bias yielding more button presses (23, 45; F = 2.5599, p <.05). Practically, as there appears not to be a relationship between button presses and patient race, we do not think this is a practically meaningful result and only a statistical finding.

Model 3 (Conditions 1V + 2R)

PERFORMANCE (P) SCORE: Table 17 shows how motivation (p = .0159) and condition (p = .0023) were significant predictors of p-score with a good model fit (F = 1.4332, p = .2517) and 21.766% of the variation explains (adjusted Rsquare, [F(14, 82) = 2.9078, p = .0013]. VIF for all parameters was < 5. Condition 1V was slower and less accurate (avg. p-score of .7867) than 2R (1.2958). More motivated participants were more accurate but not faster.

Table 17 - Model 3 1V+2R, Performance (P) Score Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	0.3825	0.3007	0.9111
IAT Categories	0.4146	0.5432	0.6541
Patient Sex	0.0329	0.1293	0.7201

Participant Age	0.0066	0.0258	0.8727
Patient Race	0.26	0.5108	0.6019
Motivation	**1.5432**	**6.0648**	**0.0159**
Condition (1V or 2R)	**2.5164**	**9.8891**	**0.0023**

ERROR DIRECTION: Triage category (p = .0040) and patient sex (p = .0002) were the significant predictors of error direction overall [see Table 18]. The model had good fit (F = 1.1091, p = .45) and explained a significant portion of the variation in error direction [F(13, 80) = 4.604, p < .0001] with an adjusted Rsquare of .337. Across both conditions, male patients were more likely to be triaged more severely than they should be while female patients were triaged less severely than they should be.

Table 18 - Model 3 1V+2R, Error Direction Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	**15.4886**	**4.1824**	**0.004**
IAT Categories	1.2523	0.4509	0.7174
Patient Sex	**13.6679**	**14.7632**	**0.0002**
Participant Age	0.361	0.3899	0.5341
Patient Race	1.9022	1.0273	0.3626
Motivation	0.1164	0.126	0.7236
Condition (1V or 2R)	1.8362	1.9834	0.1629

TIME: IAT category, motivation, and age were significant predictors of time. However, this model violated the goodness of fit test (p < .05) and thus does not fit the data well.

BUTTON PRESSES: Condition (p = .0069), motivation (p =.0008), and IAT categories (p = .0002) were significant predictors of p-score [see Table 19]. The model had good fit (F = 1.4369, p = .2489) and explained a significant portion of the variation [F(14, 81) = 4.8126, p <.0001] with an adjusted rsquare of .4149. VIF was < 5 for all parameters. Condition 1V had less button presses than 2R, lower motivation as associated with more button presses overall, and minor pro black bias yielded more button presses than no bias, and even more than negative black bias.

Table 19 - Model 3 1V+2R, Button Presses Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	21.24	0.3234	0.8975
IAT Categories	289.5004	7.3475	0.0002
Patient Sex	0.3947	0.0301	0.8628
Participant Age	10.0825	0.7677	0.3835
Patient Race	15.5204	0.5909	0.5562
Motivation	160.486	12.2193	0.0008
Condition (1V or 2R)	100.9825	7.6887	0.0069

Model 4 (Between conditions 1V and 2R) – Interaction Effects

PERFORMANCE (P-SCORE): Motivation, condition, and patient race*patient sex were significant predictors of p-score [F(20, 76) = 2.7326, p = .0009](see Table 20). There was good model fit (F = 1.3403, p = .2985) and the model had an adjusted Rsquare of .265. 1V virtual patients were triaged slower and less accurately than 2R real patient actors. Men of color were triaged less accurately and slower than any other patient (.54). White men were triaged the most accurately and quickly (1.28). Women of color were triaged more accurately (1.09) than white women (.91), though their confidence intervals have substantial overlap. Men of color had the widest confidence interval indicating a wider range of variances in their triage experience than any other group.

Table 20 - Model 4 1V+2R, Performance (P) Score Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	0.7271	0.6085	0.6936
IAT Categories	0.2092	0.2917	0.8313
Patient Sex	0.1465	0.613	0.4361
Participant Age	<0.0001	0.0001	0.9921
Patient Race	0.4826	2.0195	0.1594
Motivation	**1.2628**	**5.2838**	**0.0243**
Condition (1V or 2R)	**3.4811**	**14.5657**	**0.0003**
Race x IAT	0.5949	0.8297	0.4816
Race x Sex	**1.6161**	**6.7623**	**0.0112**
Race x Sex x IAT	1.8265	2.5475	0.0621

ERROR DIRECTION: Triage category, patient sex, patient race, and patience race*sex were significant predictors of error direction (see Table 21). The model was well fitted (F = .9624, p = .5739) and accounts for a significant [F(19, 74) = 4.4378, p <.0001] portion of the variance in error direction (Adjusted Rsquare of .41). Overall, when the category was incorrectly assigned, male patients were triaged more severely than they should be (-1.07) and female patients did not have a specific error direction (.03). Patients of color were triaged more severely than they should be (-.99) and white patients in no particular direction (-.05). Men of color were triaged much more severely than they should be (-2.09) than white men (-.05), white women (-.05), or women of color (.11).

Table 21 - Model 4 1V+2R, Error Direction Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	**15.027**	**4.5781**	**0.0023**
IAT Categories	2.4942	1.0132	0.3918
Patient Sex	**19.8617**	**24.204**	**<0.0001**

Participant Age	0.521	0.6349	0.4281
Patient Race	**5.2324**	**6.3763**	**0.0137**
Motivation	0.0175	0.0213	0.8845
Condition (1V or 2R)	2.9812	3.6329	0.0605
Race x IAT	1.7888	0.7266	0.5393
Race x Sex	**9.2524**	**11.2751**	**0.0012**
Race x Sex x IAT	4.1416	1.6824	0.1781

TIME: The model was poorly fitted for time ($F = 3.06$, $p < .05$) despite all VIFs below 5.

BUTTON PRESSES: IAT category, motivation, condition, and race*IAT score were significant predictors (see Table 22). The model had good fit ($F = 1.3429$, $p = .2971$) and accounts for a significant portion (adjusted Rsquare = .45) of the variance in button presses [$F(20, 75) = 4.9135$, $p <.0001$]. VIF scores for all parameters were < 5 indicating no collinearity problems. Patients of color received more button presses from minor positively black biased participants and less button presses than white patients from white biased responders.

Table 22- Model 4 1V+2R, Button Presses Effect Tests

Source	Sum of Squares	F-Ratio	P-Value
Triage Category	21.2485	0.3453	0.8837
IAT Categories	**368.5787**	**9.9819**	**<0.0001**
Patient Sex	5.7633	0.4682	0.4959
Participant Age	7.5541	0.6137	0.4358
Patient Race	19.5192	1.5859	0.2118
Motivation	**150.373**	**12.2173**	**0.0008**
Condition (1V or 2R)	**89.215**	**7.2484**	**0.0087**
Race x IAT	**112.6525**	**3.0509**	**0.0337**
Race x Sex	11.9826	0.9735	0.237
Race x Sex x IAT	37.8946	1.0263	0.3859

Overall, Model 1 offered a better fit for condition-based evaluation of each dependent variable. There were already some concerns of overfitting with model 1 so it is unsurprising that we had more extreme issues with poor model fit when we included interactions. It is possible that there were interaction effects of interest, but we do not have enough data to evaluate those at this time. Fortunately, we did have enough data to evaluate interaction effects on the full dataset (model 4). Overall, model 4 offered a better and more nuanced analysis than model 3.

5.4 Discussion

This work is an extension of Chapter 4 where we evaluated triage accuracy, error direction, completion times, p-scores (accuracy + time), and button presses between a virtual patient simulation (1V) and a 'real' patient actor simulation (R2). With the addition of some new variables and a different analysis approach, we now consider whether the triage category of the patient, the patient's race, the patient's gender, the participant's motivation, or the participant's age have any interactive effect with the evaluated dependent variables. Moreover, we explored whether any potential bias could be anticipated/captured by the IAT. We also explored some potential interaction effects like patient race*patient gender to see if intersectional identities had any relationship with triage performance.

5.4.1 Exploring Bias

According to the IAT scores, one participant had a negative black bias, seven had a moderate negative black bias, two had a minor negative black bias, two had no bias in either direction, and one had a minor positive black bias. When pooling together both conditions, we found that there was a significant effect of IAT on overall accuracy. Overall accuracy refers to the overall accuracy rate for a participant after an entire 1V or 2R session. Worded differently, if a participant triaged 5 patients in the entire 2R condition and got 4 of them correct, their overall accuracy for 2R is 80%. If we look at the conditions individually, we still see a significant effect of IAT on overall accuracy for 2R but not for 1V. All four of our primary models evaluated the IAT and found IAT was only a significant predictor for the number of buttons a participant pressed (stronger negative black bias scores pressed fewer buttons overall). Model 4 (including both 1V+2R) explored the interaction effect between IAT and patient race and found a significant effect of IAT*Race on button presses where patients of color received more button presses from minor positively black biased participants and less button presses than white patients from white biased responders. Otherwise, there were no statistically significant effects or relationship with IAT on race-based performance evaluations. We hypothesize this could be for a few reasons. 1) we found an effect with button presses but nothing else, so perhaps we do not have enough data for the other variables, 2) The IAT is impacted by other factors like motivation and age, 3) the IAT perhaps captured a bias but not one that could be translated to a manifestation of biased behavior, or 4) the nature of MCI triage makes it less susceptible to implicit bias (3 and 4 are discussed in more detail in 5.4.3).

It is possible that we do not have a large enough sample size (or a high enough proportion of patients of color) to see an effect on other variables. We think this small participant pool exacerbates what we noted earlier in the paper, that there are questions about whether the IAT can actually predict real-world discriminatory behavior [128]. But if the IAT cannot capture real-world discriminatory behavior, what is it

capturing? We also noted earlier that the IAT can be influenced by other factors [133]. Our earlier analysis revealed that participant motivation and age were both significant predictors of IAT scores. Older participants were significantly more motivated, more motivated participants had significantly lower IAT bias, and older participants had significantly lower IAT bias. While there seems to be a three-way relationship here, an exploratory model revealed that age and age*motivation were significant predictors of IAT scores. The lack of collinearity between these factors indicates age and motivation are still unique factors from each other. This, coupled with the lack of IAT relationship on race-based triage performance and the significant relationship between age and motivation with IAT, indicates that age, motivation, and other factors may be playing a larger role than bias itself within the IAT scores. While age had a direct significance with completion time, this effect was non-significant in the more complex models (except for the Model 1 2R and Model 2 2R models with poor fit). Motivation itself was a more significant predictor across all models and was related to response time (1V), button presses (1V), p-score (2R), p-score(1V+2R), and button presses (1V+2R). Even taking into account interaction effects, motivation was significant for p-score (1V+2R) and button presses (1V+2R). More motivated participants were faster and pressed fewer buttons, indicating that they were more engaged with the actual triage task. This could be a potential metric for evaluation of a responder's triage performance.

For the relationship between age and IAT, we hypothesize that older responders have had more experience working with diverse people of the span of their career. They have likely had more opportunities to be challenged by biases in their careers. Our least motivated participant had by far the worst bias scores, indicating perhaps IAT can be worsened by a lack of engagement. Alternatively, perhaps more experienced and motivated responders are simply offering more cognitive effort to not be biased.

Despite the IAT not bearing much association with the dependent variables or when combined with patient race, we did capture some effect of patient race when combined with patient sex on p-score and error direction (when combining the data from the virtual and real simulations together). Men of color were triaged less accurately and slower than any other demographic, while white men were triaged the most accurately and quickly. Women of color were triaged slightly more accurately than white women but the confidence intervals have so much overlap, this effect may not actually be present. Now that we know the overall accuracy, it is important to understand the types of errors being made. E.g., are women of color more likely to be under triaged, etc? When responders were incorrect, we found that male patients overall were given a higher priority status than they should have been. Comparatively, women had such a broad range of error directions that their error direction score came out close to 0.

This indicates that women receive a wider variability of care. Patients of color tended to be triaged at a higher priority (greater severity) than white patients. Men of color specifically were triaged much more severely (-2) than any other group—so far as to yield a negative or neutral effect whether they priority care. Figure 30 outlines how this works. A patient who should be in the yellow category triaged -2 will go down to 'grey'. White men and women shared an error direction of -.05, indicating a very slight inclination to triage them at higher priority. Women of color however, had a slight error direction of .11, indicating a slight tendency to triage them at a lower priority. Exploring the main effect of gender alone (regardless of race) revealed that men were more likely to be given a more severe triage category than women in both the 1V condition and the combined (full) model.

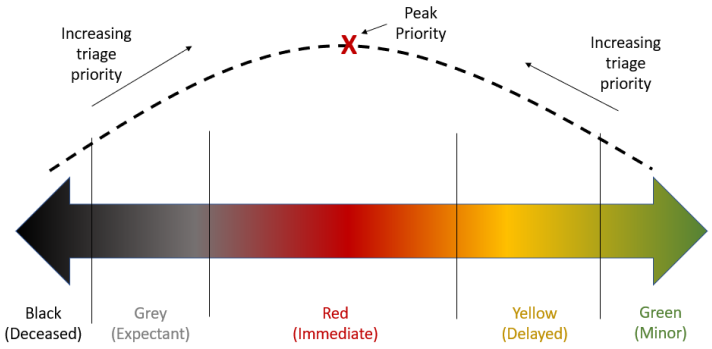

Figure 30: Triage Priority by Category

This study shows the beginning of potential gender and race-based trends in triage accuracy. Overall, we found the biggest differences between men vs women, and white men vs. men of color. Men tended to get more priority care than women, with black men getting the most severe triage categories. We found no effects of time, motivation, and only one minor relationship with IAT on button presses. This means that certain groups were not allocated more time or VTT engagement than others.

5.4.2 The Effect of Triage Category

We acknowledge that it is most difficult to accurately triage between grey/red and red/yellow. Thus, we explored whether triage category had a predictive effect on any of the dependent variables. Triage category alone had a significant effect on error direction. Responders tended to triage grey up into red and yellow down into red, indicating an inclination to 'play it safe' and round toward a higher priority in condition 2R and the full interaction model (including both 2R and 1V). Even though we have considered more factors like motivation this time, this finding is in line with the same discovery in Chapter 4.

5.4.3 Did We Actually Capture Bias?

In Chapter 4, we outlined how different the performance was in the 1V condition compared to 2R (most saliently in p-score). When we combined the two conditions together for analysis, we did find significant effects of condition on error direction, button presses, and p-score. 1V patients were triaged more severely, less accurately, and had less button presses in the VTT. Our bias findings are mostly subtle and limited to the full combined models. Thus, we acknowledge that perhaps the 1V simulation is swaying the data and

indicating a greater bias than there is in reality. Part of our issue in finding these bias effects in the 2R condition alone could be due to a lack of participants (or not enough patients of color), or it could be due to MCI triage itself.

Implicit racial bias has several hypothesized cognitive vehicles for its manifestation. Here we will mention two: Attribution bias and confirmation bias. Fundamental attribution error in healthcare refers to when healthcare providers may attribute health concerns to personal choices rather than systemic factors [134]. Confirmation bias is the cognitive tendency for people to interpret and favor information that confirms or supports their existing beliefs [135]. In healthcare, attribution bias may manifest as a lack of empathy for certain patients as the injury may be perceived as 'their fault.' Confirmation bias may cause biased interpretations of patient behavior, like believing a patient in pain is faking it (therefore not taking the pain seriously [115]). In an MCI, the cause of injuries is salient, external to the patient, and the responder is also subsumed in the same environment which gives the responder a constant external cue. It is plausible that this external cue plays a role in mitigating attribution bias, as the patient's injuries resulting from an explosion are clearly not attributable to any personal failing on their part. This external circumstance provides a clear external cause, potentially reducing the tendency to attribute the situation solely to the patient's character or behavior. Moreover, the presence of such a conspicuous external event could alleviate confirmation bias. The patient's reported pain or expressed concerns become highly salient and easily explainable within the context of the MCI. This situational clarity may prompt healthcare providers to consider the patient's symptoms and needs more objectively, focusing on the immediate and understandable impact of the incident rather than predispositions towards preconceived notions or assumptions (though this may all depend on the nature of a given MCI). Regarding the IAT, it is possible that the biases involved in healthcare (attribution and confirmation) are not well captured by the association (attitude) based measurement style of the IAT [136]. Good/bad or positive/negative words like 'laughter, horrible, glorious, and nasty' may capture attitudes underlying and informing stereotypes, but the stereotypes themselves may be a more approximate consideration for attribution and confirmation bias workings and impacts. Worded differently, attitudes may be the kindling for a stereotype to take root, but a biased attitude itself may not rise to the level of a held stereotype, which may be a prerequisite for attribution bias or confirmation bias effects.

Given the highly sensitive nature of triage, these race/gender-based differences could play a downstream role in a patient's life or death. Triage represents a significant yet fleeting interaction with patients. While its severity renders patients more vulnerable (especially to overt racism), the impacts of implicit bias in healthcare may also manifest more subtly, akin to a 'death by a thousand papercuts,' which are often smaller and challenging to quantify at the initial phase of a patient's healthcare journey. Earlier, we discussed institutional racism as the systemic embedding of discriminatory practices within healthcare systems

[9]. This systemic issue can lead to diminished healthcare outcomes for patients of color, influenced by interpersonal, organizational, and structural factors of systemic racism. Studies have shown that black patients experience worse health outcomes in emergency room settings [23], [24], which operate within distinct institutional contexts compared to the chaotic environment of a Mass Casualty Incident (MCI) field. To our knowledge, there is not a study evaluating triage in the field of the MCI. It is plausible that the negative outcomes resulting from bias are not solely attributable to individual experiences but are deeply embedded within broader healthcare structures [135]. This complexity makes it challenging to fully capture and address these issues during the initial triage phase of an MCI response. Combined with the potentially increased objectivity of MCIs, this may be why we did not capture racially biased performance.

Our final hypothesized reason why we see a mixed bias effect is that the simulation was not a live-scenario. In a live-scenario, responders' own lives can be at risk and there are copious unknowns, fostering a cognitively demanding "…breeding ground where implicit biases particularly influence clinical decisions" [8]. The pressure of a developing and impending hazard may change the way responders engage in triage [137] and could potentially cause responders to fall back on implicit biases due to exhausting cognitive load on hazard awareness. However, we may also see this mixed effect due to the risk of inflated type-I error from comparing several models.

5.5 Conclusion

This work explored whether there were measurable, race-based differences in MCI triage accuracy across two simulations (one with virtual patients and one with patient actors). We found there were some effects of triage category on the types of errors made by responders. Moreover, we found that patient race, gender, and gender*race could impact performance and types of errors beyond triage category. In our study we revealed that men of color were triaged less accurately and slower than any other demographic, while white men were triaged the most accurately and quickly. Women of color and white women were triaged somewhere in the middle. We also found that the IAT did not 'catch' any of these potentially biased behaviors because the IAT had little to no statistical relationship with any of the outcome variables. We argue that this could be due to factors like motivation and age (and perhaps expertise) interacting with the IAT, or that our simulation did not sufficiently pressure responders during the task.

5.5.1 Implications, Limitations, & Future Work

Though our data is not conclusive, we have evidence to indicate that there may be a relationship between racial bias, gender bias, and intersectionality on several metrics of triage performance. This means a few things:

1) There is potential for study stimuli to have an impact on measured outcome variables and this should be further explored,
2) future work needs to continue exploring potential racial bias in the MCI field itself,
3) future work needs to evaluate whether there is a worsening bias effect in a fully virtual condition (perhaps due to diminishing humanness [138])
4) and poor virtual conditions could obscure or sway data more than a 'real' simulation.

As there were 10 patients to triage in the 2R condition and on average only 7 were triaged, our future work will use the video data we captured to explore whether responders tended to prioritize approaching certain groups of patients.

Unfortunately, given the small sample size, there were some analyses we did not have sufficient power to explore. The small number of female participants meant we could not explore any participant-gender-based differences in triage accuracy informed by biases. Moreover, the lack of black participants meant we could not evaluate race-based differences in triage accuracy as informed by biases either. Future work should capture more gender and racially diverse participants, as well as more participants generally.

An interesting finding was that less motivated participants pressed more buttons and took longer to complete their triage tags but did not have depreciated accuracy. This could mean that the diminished motivation slows responders down and causes them to engage with the VTT more as they consider the appropriate triage tag assignment. Our future work can explore not only the number of button presses, but what specific facets of the VTT participants were engaging with. However, as we ran four models with several factors, it is possible that we have some alpha-inflation (type I error) impacting the findings outlined in this document. Future work should consider Bayesian inference statistical strategies to evaluate these analyses based on the likelihood between H0 and H1.

6 Bringing it All Together

The contributions of this research are multifaceted, addressing the design, evaluation, and impact of augmented reality (AR) tools in the context of mass casualty incident (MCI) triage.

Contribution 1 focuses on the development and implementation of an applied, iterative user-centered design (UCD) and evaluation process for an AR triage tool suite. This contribution highlights the significance of user engagement and feedback in refining AR tools for practical application, detailing AR use cases such as enhancing situational awareness, communication, and decision-making in emergency response. UCD centers on the interface between humans and technology, focusing on end users to elucidate their needs, preferences, and behaviors while considering human capability and system requirements (outlined in contribution 2). Systems that do not consider user experiences may make something conceptually meaningful for workers but not practically usable. Our UCD allowed us to develop, in partnership with emergency responders, an AR tool suite for mass casualty incident triage. Here, we learned that the AR tool suite we developed had good usability but needed further testing and iteration. This set us up for contribution 3.

Contribution 2 delves into a cognitive work analysis specifically tailored for MCI triage. This analysis scrutinizes the cognitive processes, decision-making strategies, and information flow in triage scenarios, offering valuable insights into the challenges faced by emergency responders and proposing design considerations for future AR systems to optimize triage efficiency and accuracy. The ecological perspective considers the broader context in which human activities occur, including physical, social, and organizational environments. CWA advocates for a holistic approach rooted in the belief that human behavior cannot be fully understood without considering the environment in which it takes place. A UCD focuses on a user's perspective of needs. However, in complex sociotechnical systems, there can be more to a work environment (up or down stream of them) than a user may be able to report or consider. Failure to consider broader work domain constraints could yield a 'usable' interface that does not actually help manage the work environment. Our exploration of the MCI triage work domain situated us to improve the AR tool suite and design the simulations featured in contributions 3 and 4. Contributions 1 and 2 give us a breadth of understanding not achievable with only one.

Contribution 3 outlines the development and evaluation of two progressively context-intensive simulations. These simulations, which included virtual patients and 'real' patient actors, served as a test bed for the Virtual Triage Tag (VTT) described in Contribution 1. Given the high-risk and unpredictable nature of MCI triage, it is unethical and infeasible to put the VTT in the field without further testing. However, this also means that we cannot test in the live field, hence the demand for simulations in order to evaluate the

usability of the VTT. Here, we briefly explore the success of each simulation (via accuracy). Moreover, the chapter evaluates and compares these simulations, incorporating responder feedback from semi-structured interviews to provide a comprehensive assessment of the VTT. This chapter gives us a better understanding of the VTT's effectiveness in context and provides some meaningful directions for continued iteration. Regarding the simulations, we found that the virtual simulation had poor accuracy despite it being designed with a training facility specializing in MCI triage simulation development.

Contribution 4 examines the impact of simulation stimuli on triage accuracy. This contribution analyzes how various simulation factors, such as patient race and gender, may influence decision-making and task performance, offering insights into potential factors that can affect triage performance. We find that patient race and gender were not significant predictors of performance within each simulation, though they may have slight impacts when the data from both simulations are combined.

Together, this body of work provides holistic, novel, methodological contributions to mass casualty incident triage via a thorough exploration of AR technology's potential to enhance emergency response, grounded in user-centered design, cognitive analysis, and realistic simulation testing.

- Holistic: This body of work examines MCI triage from the user perspective, work perspective, and evaluation perspective.
- Novel: To our knowledge, we are the only example of a UCD and CWA for AR in MCI triage. Moreover, we offer the first bias-based evaluation of triage performance in the 'field' (via simulation) as opposed to the emergency room.
- Methodological: We offer example applications of UCD in an academic development cycle and create a foundation for future CWA and simulation development in this space. We also provide insight into simulations for usability evaluations in high-risk environments. Finally, we explore impacts of race and gender bias on triage impacts, indicating the potential for study stimuli to impact user performance.

Altogether, this work provides a foundation for other developers whether they want to explore AR triage tools, AR simulations, or bias in healthcare.

7 References

[1] World Health Organization, "Mass Casualty Management Systems Strategies and guidelines," 2007.

[2] A. K. Donahue and R. V. Tuohy, "'Lessons we don't learn: a study of the lessons of disasters; why we repeat them; and how we can learn them," *Homel. Secur. Aff.*, vol. 2, no. 2, pp. 1–28, 2006.

[3] M. Sommer, G. S. Braut, and O. Nja, "A model for learning in emergency response work," *Int. J. Emerg. Manag.*, vol. 9, no. 2, pp. 151–169, 2013, doi: 10.1504/IJEM.2013.055161.

[4] U. Rosenthal, R. A. Boin, and L. K. Comfort, *Managing Crises: Threats, Dilemmas, Opportunities,*. Springfield, Illinois, 2001.

[5] M. D. R. F. Fernández, C. I. T. Bernabe, and R. R. Rodríguez, "Red cross triage app design for augmented reality glasses," *ACM Int. Conf. Proceeding Ser.*, vol. 03-05-Nove, no. 2901, pp. 11–14, 2014, doi: 10.1145/2676690.2676697.

[6] D. Alexander, "Towards the development of a standard in emergency planning," *Disaster Prev. Manag.*, vol. 14, no. 2, pp. 158–175, 2005.

[7] G. Kurillo, A. Y. Yang, V. Shia, A. Bair, and R. Bajcsy, "New emergency medicine paradigm via augmented telemedicine," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9740, pp. 502–511, 2016, doi: 10.1007/978-3-319-39907-2_48.

[8] A. Bagnis, E. Caffo, C. Cipolli, A. De Palma, G. Farina, and K. Mattarozzi, "Judging health care priority in emergency situations: Patient facial appearance matters," *Soc. Sci. Med.*, vol. 260, no. July, 2020, doi: 10.1016/j.socscimed.2020.113180.

[9] D. R. Williams, J. A. Lawrence, and B. A. Davis, "Racism and Health: Evidence and Needed Research," *Annu. Rev. Public Health*, vol. 40, pp. 105–125, 2019, doi: 10.1146/annurev-publhealth-040218-043750.

[10] D. A. Saucier, C. T. Miller, and N. Doucet, "Differences in helping whites and blacks: A meta-analysis," *Personal. Soc. Psychol. Rev.*, vol. 9, no. 1, pp. 2–16, 2005, doi: 10.1207/s15327957pspr0901_1.

[11] I. W. Maina, T. D. Belton, S. Ginzberg, A. Singh, and T. J. Johnson, "A decade of studying implicit racial/ethnic bias in healthcare providers using the implicit association test," *Soc. Sci. Med.*, vol. 199, pp. 219–229, 2018, doi: 10.1016/j.socscimed.2017.05.009.

[12] K. White and L. N. Borrell, "Racial/ethnic residential segregation: Framing the context of health risk and health disparities," *Heal. Place*, vol. 17, no. 2, pp. 438–448, 2011, doi: 10.1016/j.healthplace.2010.12.002.

[13] H. Landrine, I. Corral, J. G. L. Lee, J. T. Efird, M. B. Hall, and J. J. Bess, "Residential Segregation and Racial Cancer Disparities: A Systematic Review," *J. Racial Ethn. Heal. Disparities*, vol. 4, no. 6, pp. 1195–1205, 2017, doi: 10.1007/s40615-016-0326-9.

[14] M. T. Hicken, N. Kravitz-Wirtz, M. Durkee, and J. S. Jackson, "Racial inequalities in health: Framing future research," *Soc. Sci. Med.*, vol. 199, no. December 2017, pp. 11–18, 2018, doi: 10.1016/j.socscimed.2017.12.027.

[15] C. Fitzgerald and S. Hurst, "Implicit bias in healthcare professionals: A systematic review," *BMC Med. Ethics*, vol. 18, no. 1, 2017, doi: 10.1186/s12910-017-0179-8.

[16] A. G. Greenwald, N. Dasgupta, J. F. Dovidio, J. Kang, C. A. Moss-Racusin, and B. A. Teachman, "Implicit-Bias Remedies: Treating Discriminatory Bias as a Public-Health Problem," *Psychol. Sci. Public Interes.*, vol. 23, no. 1, pp. 7–40, 2022, doi: 10.1177/15291006211070781.

[17] L. A. Cooper *et al.*, "The associations of clinicians' implicit attitudes about race with medical visit communication and patient ratings of interpersonal care," *Am. J. Public Health*, vol. 102, no. 5, pp. 979–987, 2012, doi: 10.2105/AJPH.2011.300558.

[18] D. J. F, G. S. L, and K. Kerry, "Intergroup Contact: The Past, Present, and the Future," *Gr. Process. Intergr. Relations*, vol. 6, no. 1, pp. 5–21, 2003.

[19] A. R. Green *et al.*, "Implicit bias among physicians and its prediction of thrombolysis decisions for black and white patients," *J. Gen. Intern. Med.*, vol. 22, no. 9, pp. 1231–1238, 2007, doi: 10.1007/s11606-007-0258-5.

[20] J. W. Kunstman and E. A. Plant, "Racing to Help: Racial Bias in High Emergency Helping Situations," *J. Pers. Soc. Psychol.*, vol. 95, no. 6, pp. 1499–1510, 2008, doi: 10.1037/a0012822.

[21] C. Arslanian-Engoren, "Gender and age bias in triage decisions," *J. Emerg. Nurs.*, vol. 26, no. 2, pp. 117–124, 2000, doi: 10.1016/S0099-1767(00)90053-9.

[22] R. D. Axelson, C. M. Solow, K. J. Ferguson, and M. B. Cohen, "Assessing implicit gender bias in medical student performance evaluations," *Eval. Heal. Prof.*, vol. 33, no. 3, pp. 365–385, 2010, doi: 10.1177/0163278710375097.

[23] C. D. Schrader and L. M. Lewis, “Racial disparity in Emergency Department triage,” *J. Emerg. Med.*, vol. 44, no. 2, pp. 511–518, 2013, doi: 10.1016/j.jemermed.2012.05.010.

[24] C. Pollock, “A preliminary study exploring racial differences in triage, hospitalization status, and discharge medication in an emergency department in Graniteville, SC,” *ProQuest Diss. Theses*, p. 59, 2013.

[25] C. R. Nelson, J. L. Gabbard, J. B. Moats, and R. K. Mehta, “User-Centered Design and Evaluation of ARTTS: an Augmented Reality Triage Tool Suite for Mass Casualty Incidents,” *Proc. - 2022 IEEE Int. Symp. Mix. Augment. Reality, ISMAR 2022*, pp. 336–345, 2022, doi: 10.1109/ISMAR55827.2022.00049.

[26] C. Stephanidis *et al.*, “Seven HCI Grand Challenges,” *Int. J. Hum. Comput. Interact.*, vol. 35, no. 14, pp. 1229–1269, 2019, doi: 10.1080/10447318.2019.1619259.

[27] R. T. Azuma, “A survey of augmented reality,” *Found. Trends Human-Computer Interact.*, vol. 6, no. 4, pp. 335–385, 1997, doi: 10.1561/1100000049.

[28] E. Klopfer, *Augmented Learning: Research and Design of Mobile Educational Games*. 2008.

[29] J. D. Gould and C. Lewis, “Designing for usability: Key principles and what designers think,” *Commun. ACM*, vol. 28, no. 3, pp. 300–311, 1985, doi: 10.1145/3166.3170.

[30] P. S. P. Rex Hartson, “The Wheel: UX Processes, Lifecycles, Methods, and Techniques 2,” in *The UX Book: Agile UX Design or a Quality User Experience*, 2018, pp. 27–48.

[31] P. S. P. Rex Hartson, “What are UX and UX Design? 1,” in *The UX Book: Agile UX Design or a Quality User Experience*, 2018, pp. 2–25.

[32] T. C. Peck, K. A. McMullen, J. Quarles, K. Johnsen, C. Sandor, and M. Billinghurst, “DiVRsify: Break the Cycle and Develop VR for Everyone,” *IEEE Comput. Graph. Appl.*, vol. 41, no. 6, pp. 133–142, 2021, doi: 10.1109/MCG.2021.3113455.

[33] K. J. Vicente, *Chapter 1: What’s the Problem? Scope and Criteria for Success; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. Lawrence Erlbaum Associates, 1999.

[34] H. Berndt, D. Wessel, T. Mentler, and M. Herczeg, “Human-centered design of a virtual reality training simulation for mass casualty incidents,” *2018 10th Int. Conf. Virtual Worlds Games Serious Appl. VS-Games 2018 - Proc.*, 2018, doi: 10.1109/VS-Games.2018.8493427.

[35] J. Gillis, P. Calyam, O. Apperson, and S. Ahmad, “Panacea’s Cloud: Augmented reality for mass casualty disaster incident triage and co-ordination,” *2016 13th IEEE Annu. Consum. Commun. Netw. Conf. CCNC 2016*, pp. 264–265, 2016, doi: 10.1109/CCNC.2016.7444772.

[36] M. Vassell, O. Apperson, P. Calyam, J. Gillis, and S. Ahmad, “Intelligent Dashboard for augmented reality based incident command response co-ordination,” *2016 13th IEEE Annu. Consum. Commun. Netw. Conf. CCNC 2016*, pp. 976–979, 2016, doi: 10.1109/CCNC.2016.7444921.

[37] E. B. Lerner *et al.*, “Mass casualty triage: An evaluation of the science and refinement of a national guideline,” *Disaster Med. Public Health Prep.*, vol. 5, no. 2, pp. 129–137, 2011, doi: 10.1001/dmp.2011.39.

[38] J. P. Killeen, T. C. Chan, C. Buono, W. G. Griswold, and L. A. Lenert, “A Wireless First Responder Handheld Device for Rapid Triage , Patient Assessment and Documentation during Mass Casualty Incidents AMIA 2006 Symposium Proceedings Page - 429 AMIA 2006 Symposium Proceedings Page - 430,” *Symp. A Q. J. Mod. Foreign Lit.*, pp. 429–433, 2006.

[39] T. Mizumoto, S. Imazu, W. Sun, N. Shibata, and K. Yasumoto, “Emergency medical support system for visualizing locations and vital signs of patients in Mass Casualty Incident,” *2012 IEEE Int. Conf. Pervasive Comput. Commun. Work. PERCOM Work. 2012*, no. March, pp. 740–745, 2012, doi: 10.1109/PerComW.2012.6197611.

[40] N. B. F. and M. D. and H. H. in N. B. California, “START - Simple Triage And Rapid Treatment,” 1983.

[41] A. Uchiyama, T. Hirao, H. Yamaguchi, and T. Higashino, “Image sensor communication for patient ID recognition using mobile devices,” *Proc. 3rd Int. Work. Sens. Appl. Mob. Phones, PhoneSense 2012 - Conjunction with ACM SenSys 2012*, 2012, doi: 10.1145/2389148.2389154.

[42] D. Jiang *et al.*, “Hierarchical cloud-fog platform for communication in disaster incident coordination,” *Proc. - 2019 7th IEEE Int. Conf. Mob. Cloud Comput. Serv. Eng. MobileCloud 2019*, pp. 1–7, 2019, doi: 10.1109/MobileCloud.2019.00008.

[43] H. Dong, J. Schafer, and A. Ganz, “Augmented reality based mass casualty incident training system,” *2016 IEEE Symp. Technol. Homel. Secur. HST 2016*, 2016, doi: 10.1109/THS.2016.7568946.

[44] G. Paravantes, H. Leehane, D. Premarathna, C. Chung, and D. L. Kappen, “Application of augmented reality for multi-scale interactions in emergency vehicles,” *Adjun. Proc. - 10th Int. ACM Conf. Automot. User Interfaces Interact. Veh.*

Appl. AutomotiveUI 2018, pp. 221–224, 2018, doi: 10.1145/3239092.3267415.

[45] J. Hanacek, S. Ip-Jewell, M. Harney, T. Coela, S. Trivedy, and S. T. P. Bandemegala, "AvatarMEDIC: Creating a next generation solar-system-wide medical capability," *Proc. Int. Astronaut. Congr. IAC*, vol. 2020-Octob, no. 1, pp. 1–16, 2020.

[46] R. Hartson and P. Pyla, *The UX Book*, Second. Morgan Kaufmann, 2018.

[47] D. Norman, *The Design of Everyday Things*, Revised Ed. Basic Books, 2013.

[48] United States Marine Corps Field Medical Training Battalion-East, "FMSO 107," Camp Lejeune, NC, 2013.

[49] T. J. Cowper and M. E. Buerger, "Improving our view of the World: Police and Augmented Reality Technology," 2003.

[50] S. Survival Products, "Triage Tag." 2021.

[51] S. Braithwaite, A. Cooper, E. B. Lerner, J. A. March, M. Touchstone, and J. Whitney, "Model uniform core criteria for mass casualty triage," 2011.

[52] T. Coskun, A. Benzina, E. Artinger, C. Binder, and G. Klinker, "User-centered development of UI elements for selecting items on a digital map designed for heavy rugged tablet PCs in Mass Casualty Incidents," *IHI'12 - Proc. 2nd ACM SIGHIT Int. Heal. Informatics Symp.*, pp. 151–160, 2012, doi: 10.1145/2110363.2110383.

[53] A. Mitaritonna, M. J. Abasolo, and F. Montero, "An Augmented Reality-based Software Architecture to Support Military Situational Awareness," *2nd Int. Conf. Electr. Commun. Comput. Eng. ICECCE 2020*, no. June, pp. 1–6, 2020, doi: 10.1109/ICECCE49384.2020.9179187.

[54] I. L. Nunes, R. Lucas, M. Simões-Marques, and N. Correia, "Augmented Reality in Support of Disaster Response," *Adv. Intell. Syst. Comput.*, vol. 592, pp. 155–167, 2018, doi: 10.1007/978-3-319-60366-7_15.

[55] S. K. Boddhu, R. P. Dave, M. McCartney, J. A. West, and R. L. Williams, "Context-aware event detection smartphone application for first responders," *Ground/Air Multisens. Interoperability, Integr. Netw. Persistent ISR IV*, vol. 8742, no. December 2010, p. 874213, 2013, doi: 10.1117/12.2016352.

[56] Y. Itoh, T. Langlotz, D. Iwai, K. Kiyokawa, and T. Amano, "Light Attenuation Display: Subtractive See-Through Near-Eye Display via Spatial Color Filtering," *IEEE Trans. Vis. Comput. Graph.*, vol. 25, no. 5, pp. 1951–1960, 2019, doi: 10.1109/TVCG.2019.2899229.

[57] J. Nielsen and R. L. Mack, *Usability Inspection Methods*, 1st ed. Wiley, 1994.

[58] R. P. Christopher D. Wickens, Justin G. Hollands, Simon Banbury, "Attention in Perception and Display Space," in *Engineering Psychology and Human Performance 4th ed.*, 2015, pp. 49–83.

[59] J. L. Gabbard, G. M. Fitch, and H. Kim, "Behind the Glass: Driver Challenges and Opportunities for AR Automotive Applications," *Proc. IEEE*, vol. 102, no. 2, pp. 124–136, Feb. 2014, doi: 10.1109/JPROC.2013.2294642.

[60] C. Klier and L. G. Buratto, "Stress and long-term memory retrieval: A systematic review," *Trends Psychiatry Psychother.*, vol. 42, no. 3, pp. 284–291, 2020, doi: 10.1590/2237-6089-2019-0077.

[61] S. J. Robinson, J. Leach, P. J. Owen-Lynch, and S. I. Sünram-Lea, "Stress reactivity and cognitive performance in a simulated firefighting emergency," *Aviat. Sp. Environ. Med.*, vol. 84, no. 6, pp. 592–599, 2013, doi: 10.3357/ASEM.3391.2013.

[62] C. Rae Nelson *et al.*, "Exploring Augmented Reality Triage Tools to Support Mass Casualty Incidents," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 66, no. 1, pp. 1664–1666, 2022, doi: 10.1177/1071181322661337.

[63] Evacu-Aid, "Evacu-Aid Triage Tag." [Online]. Available: http://evacu-aid.com/Products.html.

[64] J. Brooke, "SUS: A 'Quick and Dirty' Usability Scale," *Usability Eval. Ind.*, no. January 1996, pp. 207–212, 2020, doi: 10.1201/9781498710411-35.

[65] S. G. Hart, "NASA-task load index (NASA-TLX); 20 years later," *Proc. Hum. Factors Ergon. Soc.*, pp. 904–908, 2006, doi: 10.1177/154193120605000909.

[66] V. B. Nikolaev and D. N. Olimpiev, "Complex analysis and evaluation of the condition of reinforced-concrete components in power-generating structures," *Power Technol. Eng.*, vol. 43, no. 5, pp. 280–286, 2009, doi: 10.1007/s10749-010-0111-6.

[67] H. L. O'Brien, P. Cairns, and M. Hall, "A practical approach to measuring user engagement with the refined user engagement scale (UES) and new UES short form," *Int. J. Hum. Comput. Stud.*, vol. 112, no. July 2017, pp. 28–39, 2018, doi: 10.1016/j.ijhcs.2018.01.004.

[68] J. H. Jr., G. T. M. Hult, C. M. Ringle, and M. Sarstedt, *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*, 3rd ed. 2021.

[69] P. J. M. Bank, M. A. Cidota, P. (Elma) W. Ouwehand, and S. G. Lukosch, "Patient-Tailored Augmented Reality Games for Assessing Upper Extremity Motor Impairments in Parkinson's Disease and Stroke," *J. Med. Syst.*, vol. 42, no. 12, pp. 1–17, 2018, doi: 10.1007/s10916-018-1100-9.

[70] K. J. Vicente, *Chapter 9: Phase 3: PR Strategies Analysis; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[71] R. P. Christopher D. Wickens, Justin G. Hollands, Simon Banbury, "Spatial Displays," in *Engineering Psychology and Human Performance 4th ed.*, 2015, pp. 84–122.

[72] R. Pramanik, "Challenges in coordination: Differences in perception of civil and military organizations by comparing international scientific literature and field experiences," *J. Risk Res.*, vol. 18, no. 7, pp. 989–1007, 2015, doi: 10.1080/13669877.2015.1043566.

[73] J. Rasmussen, A. M. Pejtersen, and L. P. Goodstein, *Cognitive System Engineering*, 1st ed. New York: John Wiley, 1994.

[74] R. M. Kilgore, O. St-cyr, and G. A. Jamieson, "From Work Domains to Worker Competencies: A Five-Phase CWA," *Appl. Cogn. Work Anal.*, pp. 29–62, 2020, doi: 10.1201/9781420063059-6.

[75] K. J. Vicente, *Chapter 2: Why Work Analysis! An Ecological Perspective; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. Lawrence Erlbaum Associates, 1999.

[76] K. J. Vicente, *Chapter 5: Toward a Formative Approach to Work Analysis: "Workers Finish the Design"; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[77] J. Gillis, P. Calyam, O. Apperson, and S. Ahmad, "Panacea's Cloud: Augmented reality for mass casualty disaster incident triage and co-ordination," *2016 13th IEEE Annu. Consum. Commun. Netw. Conf. CCNC 2016*, pp. 264–265, 2016, doi: 10.1109/CCNC.2016.7444772.

[78] S. Nestler, M. Huber, and G. Klinker, "Improving the documentation in mass casualty incidents by combining electronic and paper based approaches," *Roots Futur. Ambient Intell. - Adjun. Proceedings, 3rd Eur. Conf. Ambient Intell. AmI 2009*, no. November 2009, pp. 57–60, 2009.

[79] A. A. Cirik and D. Mendonça, "Responding to disaster in socio-technical systems: Two case studies," *Conf. Proc. - IEEE Int. Conf. Syst. Man Cybern.*, no. October, pp. 3928–3933, 2009, doi: 10.1109/ICSMC.2009.5346653.

[80] C. Bearman, S. Rainbird, B. P. Brooks, C. Owen, and S. Curnin, "A literature review of methods for providing enhanced operational oversight of teams in emergency management," *Int. J. Emerg. Manag.*, vol. 14, no. 3, pp. 254–274, 2018, doi: 10.1504/IJEM.2018.094237.

[81] P. M. Furbee *et al.*, "Realities of rural emergency medical services disaster preparedness," *Prehosp. Disaster Med.*, vol. 21, no. 2, pp. 64–70, 2006, doi: 10.1017/S1049023X0000337X.

[82] C. L. Chng, J. Collins, and S. Eaddy, "A comparison of rural and urban emergency medical system (EMS) personnel: A texas study," *Prehosp. Disaster Med.*, vol. 16, no. 3, pp. 159–165, 2001, doi: 10.1017/S1049023X00025917.

[83] S. Procee, C. Borst, R. van Paasen, and M. Mulder, "Toward Functional Augmented Reality in Marine Navigation: A Cognitive Work Analysis," *16th Int. Conf. Comput. IT Appl. Marit. Ind.*, no. c, pp. 298–312, 2017.

[84] L. N. Clark *et al.*, "Usability evaluation of an emergency department information system prototype designed using cognitive systems engineering techniques," *Appl. Ergon.*, vol. 60, pp. 356–365, 2017, doi: 10.1016/j.apergo.2016.12.018.

[85] K. M. Feigh, M. C. Dorneich, and C. C. Hayes, "Toward a characterization of adaptive systems: A framework for researchers and system designers," *Hum. Factors*, vol. 54, no. 6, pp. 1008–1024, 2012, doi: 10.1177/0018720812443983.

[86] K. J. Vicente, *Chapter 7: Phase 1: Work Domain Analysis; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[87] N. A. Stanton, P. R. G. Chambers, and J. Piggott, "Situational awareness and safety," *Saf. Sci.*, vol. 39, no. 3, pp. 189–204, 2001, doi: 10.1016/S0925-7535(01)00010-8.

[88] M. Endsley, "Design and Evaluation for Situational Awareness Enhancement," *Proc. Hum. Factors Soc. 32nd Annu. Meet.*, pp. 97–101, 1988.

[89] K. J. Vicente, *Chapter 8: Phase 2: Control Task Analysis; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[90] K. J. Vicente, *Chapter 3: Normative Approaches to Work Analysis: "The One Best Way!"; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[91] J. C. Sperandio, "The regulation of working methods as a function of work-load among air traffic controllers," *Ergonomics*, pp. 192–202, 1978, doi: 10.1080/00140137808931713.

[92] K. J. Vicente, *Chapter 10: Phase 4: Social Organizaiton and Cooperation Analysis; Cognitive Work Analysis: Toward*

Safe, Productive, and Healthy Computer-Based Work. 1999.

[93] G. I. Rochlin, T. R. La Porte, and K. H. Roberts, "The self-designing high-reliability organization: Aircraft carrier flight operations at sea," *Nav. War Coll. Rev.*, vol. 40, no. 4, pp. 76–90, 1987.

[94] K. J. Vicente, *Chapter 11: Phase 5: Worker Competencies Analysis; Cognitive Work Analysis: Toward Safe, Productive, and Healthy Computer-Based Work*. 1999.

[95] J. Rasmussen, "Skills, Rules, and KNoweldge: Signals, Signs, and Sybmols, and Other Distinctions in Human Performance Models," *IEEE Trans. Syst. Man. Cybern.*, vol. 13, pp. 257–266, 1983.

[96] E. W. Lincoln and M. K. Strecker-McGraw, "EMS, Incident Command," *StatPearls*, pp. 1–5, 2018.

[97] N. K. Wehbi *et al.*, "A needs assessment for simulation-based training of emergency medical providers in Nebraska, USA," *Adv. Simul.*, vol. 3, no. 1, pp. 1–9, 2018, doi: 10.1186/s41077-018-0081-6.

[98] L. Otero-Varela *et al.*, "Extended reality as a training method for medical first responders in mass casualty incidents: A protocol for a systematic review," *PLoS One*, vol. 18, no. 3 March, pp. 1–6, 2023, doi: 10.1371/journal.pone.0282698.

[99] O. Zechner *et al.*, "NextGen Training for Medical First Responders: Advancing Mass-Casualty Incident Preparedness through Mixed Reality Technology," *Multimodal Technol. Interact.*, vol. 7, no. 12, 2023, doi: 10.3390/mti7120113.

[100] L. Otero-Varela *et al.*, "Extended reality as a training method for medical first responders in mass casualty incidents: A protocol for a systematic review," *PLoS One*, vol. 18, no. 3 March, pp. 1–8, 2023, doi: 10.1371/journal.pone.0282698.

[101] H. Mitsuhara, C. Tanimura, J. Nemoto, and M. Shishibori, "Expressing disaster situations for evacuation training using markerless augmented reality," *Procedia Comput. Sci.*, vol. 192, no. 2019, pp. 2105–2114, 2021, doi: 10.1016/j.procs.2021.08.218.

[102] H. Dong, J. Schafer, and A. Ganz, "Augmented reality based mass casualty incident training system," *2016 IEEE Symp. Technol. Homel. Secur. HST 2016*, pp. 1–4, 2016, doi: 10.1109/THS.2016.7568946.

[103] N. E. Kman, J. McGrath, A. R. Panchal, M. Malone, T. Sharkey-Toppen, and D. P. Way, "Implementation of a workshop for mass casualty incident triage training using an immersive virtual reality simulation," *AEM Educ. Train.*, vol. 8, no. S1, pp. S70–S75, 2024, doi: 10.1002/aet2.10939.

[104] A. Karampourian, B. Imani, and M. Khatiban, "Comparing the Training of the Mass Casualty Incident Triage by the Virtual Workshop and Non-virtual Workshop (Face-to-face) on the Nursing Students' Learning and Satisfaction," *Heal. Emergencies Disasters Q.*, vol. 8, no. 4, pp. 261–268, 2023, doi: 10.32598/hdq.8.4.494.2.

[105] M. K. Rivard, R. E. Cash, C. B. Mercer, K. Chrzan, and A. R. Panchal, "Demography of the National Emergency Medical Services Workforce: A Description of Those Providing Patient Care in the Prehospital Setting," *Prehospital Emerg. Care*, vol. 25, no. 2, pp. 213–220, 2021, doi: 10.1080/10903127.2020.1737282.

[106] S. Dwivedi *et al.*, "Comparing the efficacy of AR-based training with video-based training," *Proc. Hum. Factors Ergon. Soc. Annu. Meet.*, vol. 66, no. 1, pp. 1862–1866, 2022, doi: 10.1177/1071181322661289.

[107] S. Vyas *et al.*, "Adaptive Training on Basic AR Interactions: Bi-Variate Metrics and Neuroergonomic Evaluation Paradigms," *Int. J. Hum. Comput. Interact.*, 2023, doi: 10.1080/10447318.2023.2250578.

[108] B. M. Byrne, "Structural Equation Modeling With AMOS," *Struct. Equ. Model. With AMOS*, no. July, 2013, doi: 10.4324/9781410600219.

[109] X. Bao, T. Gao, J. Du, and L. R. Dai, "An investigation of high-resolution modeling units of deep neural networks for acoustic scene classification," *Proc. Int. Jt. Conf. Neural Networks*, vol. 2017-May, no. 1, pp. 3028–3035, 2017, doi: 10.1109/IJCNN.2017.7966232.

[110] Y. Qin, E. Bloomquist, T. Bulbul, and J. Gabbard, "Measuring the Impacts of Ar Hmd on Users' Situation Awareness During Wood Frame Assembly Tasks," *J. Inf. Technol. Constr.*, vol. 28, no. July 2022, pp. 70–86, 2023, doi: 10.36680/j.itcon.2023.004.

[111] J. Van Dam, A. Krasner, and J. L. Gabbard, "Augmented Reality for Infrastructure Inspection with Semi-autonomous Aerial Systems: An Examination of User Performance, Workload, and System Trust," *Proc. - 2020 IEEE Conf. Virtual Real. 3D User Interfaces, VRW 2020*, pp. 743–744, 2020, doi: 10.1109/VRW50115.2020.00222.

[112] N. de O. Faria *et al.*, "The Effect of Augmented Reality Cues on Glance Behavior and Driver-Initiated Takeover on SAE Level 2 Automated-Driving," *Proc. Hum. Factors Ergon. Soc.*, vol. 65, no. 1, pp. 1342–1346, 2021, doi: 10.1177/1071181321651004.

[113] C. R. Nelson and J. L. Gabbard, "Pedagogical Design Considerations for Mobile Augmented Reality Serious Games (MARSGs): A Literature Review," *Electron. - Spec. Issue Serious Games XR*, vol. 12, no. 21, 2023.

[114] C. R. Nelson and J. L. Gabbard, "Augmented Reality Rehabilitative and Exercise Games (ARREGs): A Systematic

Review and Future Considerations," in *IEEE International Symposium on Mixed and Augmented Reality*, 2023, doi: 10.1109/ISMAR59233.2023.00118.

[115] K. M. Hoffman, S. Trawalter, J. R. Axt, and M. N. Oliver, "Racial bias in pain assessment and treatment recommendations, and false beliefs about biological differences between blacks and whites," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 113, no. 16, pp. 4296–4301, 2016, doi: 10.1073/pnas.1516047113.

[116] T. C. Peck, J. J. Good, A. Erickson, I. Bynum, and G. Bruder, "Effects of Transparency on Perceived Humanness: Implications for Rendering Skin Tones Using Optical See-Through Displays," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 5, pp. 2179–2189, 2022, doi: 10.1109/TVCG.2022.3150521.

[117] T. C. Peck, S. Seinfeld, S. M. Aglioti, and M. Slater, "Putting yourself in the skin of a black avatar reduces implicit racial bias," *Conscious. Cogn.*, vol. 22, no. 3, pp. 779–787, 2013, doi: 10.1016/j.concog.2013.04.016.

[118] E. Blanpied, J. Good, and T. Peck, "Shooter bias and socioeconomic status in virtual reality," *26th IEEE Conf. Virtual Real. 3D User Interfaces, VR 2019 - Proc.*, pp. 856–857, 2019, doi: 10.1109/VR.2019.8797679.

[119] D. R. Williams, "Racial/ethnic variations in women's health: The social embeddedness of health," *Am. J. Public Health*, vol. 92, no. 4, pp. 588–597, 2002, doi: 10.2105/AJPH.92.4.588.

[120] L. Zhang, E. A. R. Losin, Y. K. Ashar, L. Koban, and T. D. Wager, "Gender Biases in Estimation of Others' Pain," *J. Pain*, vol. 22, no. 9, pp. 1048–1059, 2021, doi: 10.1016/j.jpain.2021.03.001.

[121] A. G. Greenwald, D. E. Mcghee, and J. L. K. Schwartz, "Pleasant Vs Unpleasant Associations.," *J. Pers. Soc. Psychol.*, vol. 74, no. 6, pp. 1464–1480, 1998.

[122] S. Teige-Macigemba, K. C. Klauer, and J. W. Sherman, "A practical guide to implicit association tests and related tasks," in *Handbook of implicit social cognition: Measurement, theory, and applications*, The Guilford Press, 2010, pp. 117–139.

[123] A. G. Greenwald and B. A. Nosek, "Health of the implicit association test at age 3," *Exp. Psychol.*, vol. 48, no. 2, pp. 85–93, 2001, doi: 10.1026//0949-3946.48.2.85.

[124] A. Greenwald *et al.*, "The Implicit Association Test at age 20: What is known and what is not known about implicit bias," *Annu. Rev. Psychol.*, vol. 71, pp. 419–445, 2020.

[125] A. G. Greenwald, B. A. Nosek, and M. R. Banaji, "Understanding and Using the Implicit Association Test: I. An Improved Scoring Algorithm," *J. Pers. Soc. Psychol.*, vol. 85, no. 2, pp. 197–216, 2003, doi: 10.1037/0022-3514.85.2.197.

[126] A. G. Greenwald *et al.*, "Best research practices for using the Implicit Association Test," *Behav. Res. Methods*, vol. 54, no. 3, pp. 1161–1180, 2022, doi: 10.3758/s13428-021-01624-3.

[127] K. Fiedler, C. Messner, and M. Bluemke, "Unresolved problems with the 'I', the 'A', and the 'T': A logical and psychometric critique of the Implicit Association Test (IAT)," *Eur. Rev. Soc. Psychol.*, vol. 17, no. 1, pp. 74–147, 2006, doi: 10.1080/10463280600681248.

[128] U. Schimmack, "Invalid Claims About the Validity of Implicit Association Tests by Prisoners of the Implicit Social-Cognition Paradigm," *Perspect. Psychol. Sci.*, vol. 16, no. 2, pp. 435–442, 2021, doi: 10.1177/1745691621991860.

[129] J. T. Jost, "The IAT Is Dead, Long Live the IAT: Context-Sensitive Measures of Implicit Attitudes Are Indispensable to Social and Political Psychology," *Curr. Dir. Psychol. Sci.*, vol. 28, no. 1, pp. 10–19, 2019, doi: 10.1177/0963721418797309.

[130] A. W. Meade, "FreeIAT: An Open Source Program to administer the implicit association test.," *Applies Psychol. Meas.*, vol. 33, no. 643, 2009.

[131] A. W. Meade, "Free IAT: How it Works." [Online]. Available: https://meade.wordpress.ncsu.edu/freeiat-home/freeiat-how-it-works/.

[132] A. W. Meade, "Free IAT: FAQ and Issues." [Online]. Available: https://meade.wordpress.ncsu.edu/freeiat-home/freeiat-faq-and-issues/.

[133] M. Kappi and B. S. Biradar, "Quantifying the influence of Indian optics research: An index based on three citation indicators," *Iberoam. J. Sci. Meas. Commun.*, vol. 3, no. 1, pp. 1–16, 2023, doi: 10.47909/ijsmc.39.

[134] D. J. Beal, J. B. Ruscher, and S. B. Schnake, "No benefit of the doubt: Intergroup bias in understanding causal explanation," *Br. J. Soc. Psychol.*, vol. 40, no. 4, pp. 531–543, 2001, doi: 10.1348/014466601164966.

[135] J. and Y. B. de Houwer, "Bias in Implicit Measures as Instances of Biased Behavior under Suboptimal Conditions in the Laboratory," *Psychol. Inq.*, pp. 173–176, 2022.

[136] K. N. Morehouse and M. R. Banaji, "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," *Daedalus*, vol. 153, no. 1, pp. 21–50, 2024, doi: 10.1162/daed_a_02047.

[137] F. Alruqi, E. K. Aglago, E. Cole, and K. Brohi, "Factors Associated with Delayed Pre-Hospital Times during Trauma-

Related Mass Casualty Incidents: A Systematic Review," *Disaster Med. Public Health Prep.*, vol. 17, no. 2, pp. 1–2, 2023, doi: 10.1017/dmp.2023.187.

[138] A. Erickson, G. Bruder, G. Welch, I. Bynum, T. Peck, and J. Good, "Perceived Humanness Bias in Additive Light Model Displays," no. December, 2022.

8 **Appendix**

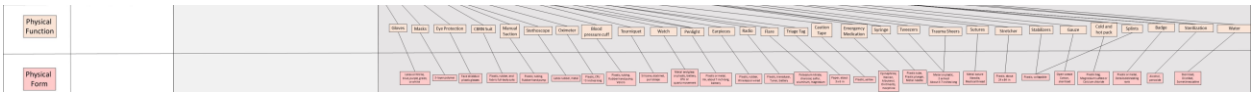

Figure 31 Remaining Elements of the ADS in Figure 7: First 3 levels of the ADS for Triage