2022

A Design Thinking Framework for Human-Centric Explainable Artificial Intelligence in Time-Critical Systems

Paul Benjamin Stone
Wright State University

A Design Thinking Framework for Human-Centric Explainable Artificial Intelligence in Time-Critical Systems

A DISSERTATION SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

by

PAUL BENJAMIN STONE

M.S.I.H.E., Wright State University, 2019

B.E., University of Huddersfield, England, 1999

2022

Wright State University

WRIGHT STATE UNIVERSITY
GRADUATE SCHOOL

12/01/2022

I HEREBY RECOMMEND THAT THE DISSERTATION PREPARED UNDER MY SUPERVISION BY PAUL BENJAMIN STONE ENTITLED A DESIGN THINKING FRAMEWORK FOR HUMAN-CENTRIC EXPLAINABLE ARTIFICIAL INTELLIGENCE IN TIME-CRITICAL SYSTEMS BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY.

Subhashini Ganapathy, Ph.D.
Dissertation Director

Ahsan Mian, Ph.D.
Program Director, Engineering Ph.D.

Shu Schiller, Ph.D.
Interim Dean of the Graduate School

Committee on Final Examination:

Subhashini Ganapathy, Ph.D.

Victor Middleton, Ph.D.

Assaf Harel, Ph.D.

Sherif Elbasiouny Ph.D.

Abstract

Stone, Paul Benjamin, Ph.D. Engineering Ph.D. Program, Department of Biomedical, Industrial, and Human Factors Engineering, Wright State University 2022. A Design Thinking Framework for Human-Centric Explainable Artificial Intelligence in Time-Critical Systems

Artificial Intelligence (AI) has seen a surge in popularity as increased computing power has made it more viable and useful. The increasing complexity of AI, however, leads to can lead to difficulty in understanding or interpreting the results of AI procedures, which can then lead to incorrect predictions, classifications, or analysis of outcomes. The result of these problems can be over-reliance on AI, under-reliance on AI, or simply confusion as to what the results mean. Additionally, the complexity of AI models can obscure the algorithmic, data and design biases to which all models are subject, which may exacerbate negative outcomes, particularly with respect to minority populations.

Explainable AI (XAI) aims to mitigate these problems by providing information on the intent, performance, and reasoning process of the AI. Where time or cognitive resources are limited, the burden of additional information can negatively impact performance. Ensuring XAI information is intuitive and relevant allows the user to quickly calibrate their trust in the AI, in turn improving trust in suggested task alternatives, reducing workload and improving task performance. This study details a structured approach to the development of XAI in time-critical systems based on a design thinking framework that preserves the agile, fast-iterative approach characteristic of design thinking and augments it with practical tools and guides. The framework establishes a focus on shared situational perspective, and the deep understanding of both users and the AI in the empathy phase, provides a model with seven XAI levels and corresponding solution themes, and defines objective, physiological metrics for concurrent assessment of trust and workload.

Executive Summary

Although there are huge benefits associated with the increased use of AI, there are also significant drawbacks such as algorithmic bias (Garcia, 2016), over-reliance (Kim et al., 2020) and automation surprise (Parasuraman & Riley, 1997). Explainable AI offers a means to mitigate these concerns (Chen et al. 2014; Ezer, 2017), but requires additional information regarding rationale, underlying biases, or potential for error to be communicated to the user (Helldin, 2014). Ensuring information is useful to the end user and improves task performance is key to successful design of explainable AI. The aim of this research is to develop a framework, based on design thinking principles, for explainable AI designs that includes evidence-based tools, approaches, and assessment methods for designers to enable explainable AI to improve user performance for time-critical systems.

There are four chapters in this dissertation. The first three chapters cover specific stages of the research with associated papers presented as published along with segue sections to provide context to the reader. Conclusions, discussion and the significance of the research are detailed in chapter four, which is described for completeness but not covered in this document as it is dependent on completion of the first three chapters.

Chapter 1 covers the exploration of the problem, developing an understanding of high-workload or time-critical AI-based decision support system. The key outputs of section are – a) a definition of the requirements for an AI-based decision support system:

Stone P.B., Ganapathy. S., (2022), submitted to the International Journal of Human-computer interaction, under review.

and b) an AI model associated with a time-critical decision support system:

Stone, P. B., Nelson, H. M., Fendley, M. E., & Ganapathy, S. (2021). Development of a novel hybrid cognitive model validation framework for implementation under COVID-19 restrictions. Human Factors and Ergonomics in Manufacturing & Service Industries, 31(4), 360-374.

Chapter 2 Expands the understanding of the problem from chapter 1, and a details structured approach for the implementation of explainable AI in time-critical systems. This section examines the nature of explainable AI, workload and trust in more detail, and defines the constructs that underpin explainable AI. The output is a design framework for the integration of explainable AI in a time-critical tasks, based on design thinking.

Stone, P. B, Jessup, S. A., Ganapathy, S., Harel, A. (2022). Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support International Journal of Human-Computer Interaction (IJHCI) Special Issue on Transparent Human-Agent Communications.

Chapter 3 focuses on the demonstration of the effectiveness of the design framework for the integration of in human-machine teaming. This is achieved through the integration of explainable AI into a trust-based tactical decision game based on drone targeting using the framework developed in chapter 2 and a human-subjects assessment of the. user performance in versions of the game with various implementations of explainable AI. The outputs of this experimental phase are a validation of the design thinking framework for explainable AI integration, evidence-based measures for task performance and a quantification of the relationships between task measures.

Chapter 4 presents the specific contributions and significance of this research.

This research contributes to the understanding of the requirements for successful implementation of explainable AI with respect to user interfaces and task performance. In addition to enhanced understanding of the problem, specific outputs of this research are the definition of user requirements for explainable AI in time-critical systems, a design framework for the implementation of explainable AI in time-critical tasks, definition of assessment metrics and success criteria for time-critical a mapping of the relationships between performance constructs including trust, workload, SA, and task performance. The significance of this contribution is to ensure that explainable AI does not come at the cost of task performance and that unintended consequences can be better predicted and avoided.

TABLE OF CONTENTS

Introduction **1**

Evolution of Research **1**

Nature and Significance of the Problem **2**

Aims and Objectives **5**

Chapter 1 – Understanding AI in Human-Machine Teams **11**

Paper 1: Development of a Novel Hybrid Cognitive Model Validation Framework for Implementation Under COVID-19 Restrictions **13**

Abstract **14**

Introduction **15**

Aims and Objectives **19**

Hybrid Cognitive Validation Framework **27**

Cognitive Model for Case Study **34**

Results **38**

Discussion and Conclusion **45**

Development of a catheterization system AI **51**

Paper 2: Deep Learning for Classification of Wall Hits in Cardiac Catheterization Simulators **54**

Abstract **55**

Background **56**

Materials and Methods **62**

Results **70**

Discussion and Conclusion **71**

Chapter 2 – An Initial Design Thinking Framework for XAI in TCS **75**

Explainable Artificial Intelligence **75**

Structured Design Methods **76**

Situational Awareness **78**

Time Critical Tasks **81**

Paper 3: Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support **83**

Abstract **84**

Introduction **85**

Objective **97**

Framework Development **98**

Framework Implementation **113**

Discussion **124**

Conclusion **127**

Chapter 3 – A Complete Design Framework for XAI in TCS **129**

Paper 4: Eye Tracking for Explainable AI (XAI) performance Assessment in Time-Critical Systems. **132**

Abstract **133**

Introduction **134**

Significance of the Research **142**

Aims **144**

Physiological Measures **145**

Assessment Platform Development **146**

Method **154**

Results **170**

Discussion and Conclusions **176**

Chapter 4 – Contributions and Future work **183**

References **190**

Appendix A - **217**

Appendix B **218**

Appendix C **219**

Appendix D **220**

Vita **221**

List of Figures

Figure 0.1	XAI Design Framework Rationale	**5**
Figure 1A.1	Cognitive Model of Cardiovascular Surgery	**37**
Figure 1B.1	Wright State University Low Cost Catheterization Simulation	**63**
Figure 1B.2	Field of Regard from Curvature Navigation Task	**65**
Figure 1B.3	Field of Regard from Bifurcation Navigation Task	**65**
Figure 1B.4	Coronary Catheters used in study	**66**
Figure 1B.5	ROC Curve for the Wall-Hit Classifier	**71**
Figure 2.1	A Model of Situated Cognition	**79**
Figure 2.2	A Dynamic Model of Situated Cognition	**80**
Figure 2A.1	Trust Process Model	**91**
Figure 2A.2	Five Stage Design Thinking Process	**99**
Figure 2A.3	Design Thinking XAI Integration Process	**112**
Figure 2A.4	Baseline Transfer of Care DSS – Patient Status Input	**117**
Figure 2A.5	Baseline Transfer of Care DSS – Patient Status Tracker	**118**
Figure2A.6	Baseline Transfer of Care DSS – Input with XAI	**124**
Figure2A.7	Baseline Transfer of Care DSS – Tracker with XAI	**124**
Figure3.1	Overview of the Design Thinking Framework (Stage 2)	**130**
Figure3.2	Stage 3 XAI Design Thinking Framework concept.	**131**
Figure3A.1	Baseline 'Captcha' Task Design.	**150**
Figure3A.2	Baseline 'Spot the Difference' Task Design	**151**
Figure3A.3	Example of the Rationale XAI Information Task	**164**

Figure3A.4 Example of Certainty XAI Information Task **165**

Figure 3A.5 Example of Rationale and Certainty XAI Information **166**

Figure 3A.6 Trust Calibration Outcome by XAI Level **174**

Figure 3A.7 Trust Calibration by XAI Level (Accurate) **174**

Figure 3A.8 Trust Calibration by XAI Level (False Negative) **175**

Figure 3A.9 Trust Calibration by XAI Level (False Negative) **175**

Figure 3A.10 Trust Calibration by Designed AI Accuracy **176**

Figure 4.1 Design Thinking Framework for XAI in TCS **183**

List of Tables

Table 0.1 Overview of Stage 1 Aims, Objectives, Research Questions & Tasks **7**

Table 0.2 Overview of Stage 2 Aims, Objectives, Research Questions & Tasks **8**

Table 0.3 Overview of Stage 3 Aims, Objectives, Research Questions & Tasks **9**

Table 1A.1 Hybrid Cognitive Validation Framework **30**

Table 1A.2 Validation Score Interpretation Framework **39**

Table 1A.3 Cognitive Task Matrix for Generic Cardiovascular Surgery Scenarios **40**

Table 1A.4 Case Study Validation Scenario Definition **42**

Table 1A.5 Case Study Validation Scenario Walkthrough and Results **42**

Table 1A.6 Hybrid Cognitive Validation Framework Score Summary **43**

Table 1A.7 Hybrid Cognitive Validation Framework Score Summary Including Objectivity Rating **44**

Table 1B.1 Confusion Matrix of catheterization Image Classifier **70**

Table 2A.1 Design Framework Task Summary **100**

Table 2A.2 Transparency Solution Themes, Properties, and Potential Impacts **107**

Table 2A.3 Hierarchical Task Analysis of Transfer of Care Task **115**

Table 3A.1 Experimental Matrix for the XAI assessment study **156**

Table 3A.2 Low Reliability (50%) Counterbalanced Experimental Design **157**

Table 3A.3 High Reliability (87.5%) Counterbalanced Experimental Design **158**

List of Abbreviations

Abbreviation	**Definition**
AI	Artificial Intelligence
APA	American Psychology Association
CNN	Convolutional Neural Network
COCATS-4	Core Cardiovascular Training Statement - 4
COVID-19	Corona Virus Disease 2019
DSS	Decision Support System
DFMEA	Design Failure Mode Effect Analysis
EU	European Union
EMS	Emergency Medical Service
EMT	Emergency Medical Technician
FMEA	Failure Mode Effect Analysis
FoR	Field of Regard
FP	False Positive
FN	False Negative
GRAD CAM	Gradient-weighted Class Activation Mapping
HCOGVM	Hybrid Cognitive Validation Model
HMT	Human-Machine Team
HTA	Hierarchical Task Analysis
IC3ST	Imperial College Complex Cannulation Scoring Tool
IUA	Interpretive Use Argument
JCS	Joint Cognitive Systems
ML	Machine Learning
MLP	Multi-Layer Perceptron
MRI	Magnetic Resonance Imaging
NEMSIS	National Emergency System Information System
NCMR	National Center for Medical Readiness
PPE	Personal Protective Equipment
PPV	Positive Predictive Value
RCA	Right Coronary Artery

RGB	Red, Green, Blue
ROC	Receiver Operating Characteristic
RNN	Recurrent Neural Network
RPD	Recognition Primed Decision
SA	Situational Awareness
SAT	SA-based Agent Transparency
SARS COV-2	Severe Acute Respiratory Syndrome Coronavirus 2
SGD	Stochastic Gradient Descent
SHAP	Shapley Additive Explanations
SME	Subject Matter Expert
SOP	Standard Operating Procedure
TCS	Time Critical System
TEXAS	Time-Critical Explainable AI System
TLX	Task Load Index
ToC	Transfer of Care
VF	Validation Framework
VGG-16	Visual Geometry Group - 2016
WHO	World Health Organization
XAI	Explainable AI
XRF	X-Ray Fluoroscopy

Glossary of Terms

Term **Definition**

Automation

“The execution by a machine agent (usually a computer) of a function that was previously carried out by a human.” (Parasuraman & Riley, 1997, p.2). For the purposes of this study, automation may or may not include an AI-based machine agent.

Artificial Intelligence

"…Is the science and engineering of making intelligent machines, especially intelligent computer programs” (McCarthy, 2007).This study considers machine learning to be a subset of Artificial Intelligence, and, in-turn, deep earning to be a subset of Machine Learning.

Deep Learning

Deep learning is defined as a representation-learning method with multiple levels of representation obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level (LeCun, Bengio, & Hinton, 2015). The structure of the deep learning network is a feed-forward system that takes an input in the form of a tensor and applies a series of hidden layers made up of one or more neurons each with a weight and activation function defined within it. (Goodfellow, Bengio, & Courville, 2016)

Explainable AI

Explainable AI is the provision of additional information to the user regarding rationale, underlying biases, or potential error of an AI (Helldin, 2014)

Intelligence

“The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets ‘intelligence’ as a form of ‘relative rationality” (Wang, 2008, p.373).

Interpretability

The ease with which an abstract concept can be readily made sense of by humans (Montavon, Samek & Muller, 2018)

Machine Learning

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy. (IBM, 2021)

Transparency

“Transparency is the quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process” (Chen et al., 2014, p. 2).

Acknowledgements

Acknowledgements of the assistance of others are always important but particularly for this study, which I completed during the COVID-19 pandemic, making the support and help of my supervisor, committee, friends, family, and colleagues, even more vital.

I would like to thank my supervisor, Dr. Subhashini Ganapathy, for her invaluable support and guidance throughout this research, particularly in the conceptualization and delivery of this dissertation. Her knowledge and experience have encouraged me throughout my academic career. I would also like to thank my committee for their advice and support throughout my research and especially in their considered responses to my defense.

I would like to thank Julie Anderson for her support and understanding during the completion of this study. I couldn’t have done it without you! Finally, I would like to thank my parents and family for all they have done for me; I would not be here today without you all and I appreciate this every day.

Make tiny changes to earth.

—Scott Hutchison

Introduction

Evolution of the research

As with most research, the problem I address in this dissertation was born out of a real-world problem. In this case, the concept was developed while I was working on the development of a catheterization training and decision support system, developing an understanding of the importance of considering such human-machine teams as joint cognitive systems (JCS). Considering human-machine systems as a unified JCS, as opposed to a stand-alone machine-based system, enables enhanced understanding of the goals and responsibilities of the system (Hollnagel and Woods, 1983). The implementation of a JCS aims to improve collaboration and performance by reducing errors due to poor communication and task assignment that can result from considering the tasks conducted by the DSS and the human agent separately (Woods 1985). More than this, a JCS is implicit in the design of a successful machine expert particularly one that relies on Artificial Intelligence (AI) and interpretation of user actions to provide adaptive decision support. More than this, a JCS is implicit in the design of a successful machine expert particularly one that relies on Augmented Intelligence and interpretation of user actions to provide adaptive Decision Support. From this starting point, I developed an understanding of the importance of considering both human and machine agents in the design of expert systems, and as my research progressed, my focus shifted from the development of expert systems for supporting cardiac catheterization to the broader problem of the design and development of expert systems, and particularly the integration of AI in the DSS, emphasizing the importance of human-machine teaming in this context. In these initial phases of the research, I studied the potential for AI support in the catheterization problem and developed an understanding of expert systems in the JCS context. I focused on the application of Deep Learning to catheter performance assessment and development of a cognitive model of cardiac catheterization. Here, I realized the nature of expert decision making was an important consideration, implementing the Recognition-Primed Decision (RPD) model, Klein (1993) as a key component of the cognitive model of cardiac catheterization. The RPD model is a representation of how experts make quick decisions in complex situations, particularly in complex domains, highlighting the significance of time-criticality as a driver in expert systems. Considering time-criticality as the context in the design of human-AI systems, allowed for a more general approach than the design of expert systems, which would necessarily have a narrower focus on a specific application. This led to the core problem I address in this dissertation – the development of a design framework for the integration of XAI in time-critical systems, which forms the key output of this research.

Nature and significance of the problem

As the use of AI increases, more and more systems rely on it as the only means to analyze the vast amounts of data being produced by today's systems (Martinez, 2019). Advances in AI are enabling systems that aim to improve human performance and enable operators to take on previously impossible or time consuming tasks (Ertel, 2018).

In this context, HAI teaming is increasingly important and traditionally accepted roles of human and machine agents are changing (Mercado et al., 2016). Along with the integration of AI into our lives and workplaces, there are widespread fears on the dangers of AI (Martinez et al., 2019). The importance of the explainable AI (XAI) problem was highlighted by Ezer et al. (2019) in their Trust engineering challenge. This research challenge included topics on XAI and Algorithmic Transparency but also areas such as user interfaces, information modalities and communication strategies, all of which intersect in this research proposal.

Explainable AI is itself a complex, multifaceted problem (Helldin, 2014) covering problems as diverse as data bias (Datta et al., 2015; Johnson, 2020), explainable algorithms in AI (Kim et al., 2020), ethics and regulation (EU, 2018; Kratsios, 2018; Larsson & Heintz, 2020) and the open sharing and reproducibility of results (Haibe-Kains, et al., 2020). Explainable AI and Transparency are often used interchangeably, but here, transparency is just one aspect of XAI. When considering the problem of designing human-centric XAI, constructs such as Trust, Workload, Situational Awareness (SA) and task performance require consideration and each of these problems is complex, interrelated and can vary from task-to-task, especially in time-critical systems (TCS) (Wachter et al., 2017). While there are numerous studies looking at the need for XAI and the development of technical solutions such as SHAP and GRAD CAM, there is a notable gap in the understanding of the implementation of these solutions. That is to say, which solutions are appropriate for a given task and how and when to implement them when designing user interfaces.

Transparency is a key element of explainable AI, but as yet there is not a single agreed definition, therefore, to ensure clarity and consistency throughout the research, the following definition was adopted: “Transparency is the quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process” (Chen et al., 2014, p. 2). Transparency is key to establishing trust in both human and human-machine teams, to prevent negative outcomes and create accountable systems (de Fine Licht & de Fine Licht, 2020). Transparency offers a means to ensure systems remain human-centric and empower operators and is a key aspect of building trust in both human-human and human-AI teams (de Fine Licht & de Fine Licht, 2020; Mercado et al., 2016). Through XAI, the benefits of the system can be maximized while enabling human operators to mitigate the potential unforeseen negative impacts that might arise. In TCS, there is also the potential for additional transparency information to negatively impact task performance, especially if information is poorly presented. Providing designers with an enhanced understanding of XAI integration, along with tools to guide the process can ensure effective human-machine teaming without compromising task performance.

There is a clear opportunity for AI to improve human-machine teams (HMTs) in TCS, where the power to quickly parse large amounts of data has huge potential benefits (Wachter et al., 2017) and understanding the problems of implementation in user interfaces is a key element of the overall problem (Ezer et al., 2019). The significance of this contribution is to ensure that XAI does not come at the cost of task performance and that unintended consequences can be better predicted and avoided. This understanding has implications for the integration of AI into almost any decision support or supervisory control application. The ability for operators or human team members to know when to trust automation or AI-based team members, and to enable accurate trust calibration.

A Framework for the integration of XAI in TCS will offer the means for user interface designers to better understand the implementation of explainable AI and to ensure unintended consequences can be better predicted and mitigated. Ultimately, the aim should be to improve task performance an enable designers to measure task performance more reliably. The expected direction of system performance improvement relating to integration of AI and XAI is summarized in Figure 0.1

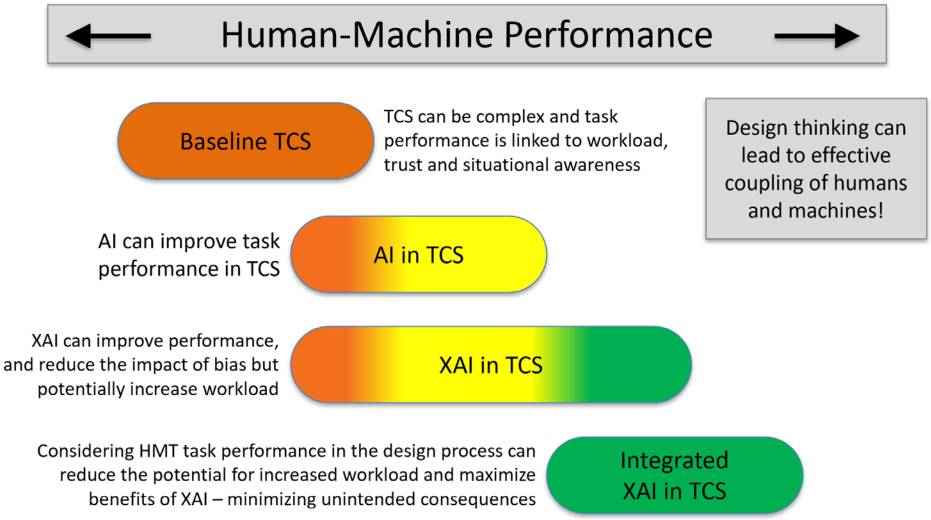

Human-Machine Performance
Baseline TCS
TCS can be complex and task performance is linked to workload, trust and situational awareness
Design thinking can lead to effective coupling of humans and machines!
AI can improve task performance in TCS
AI in TCS
XAI can improve performance, and reduce the impact of bias but potentially increase workload
XAI in TCS
Considering HMT task performance in the design process can reduce the potential for increased workload and maximize benefits of XAI – minimizing unintended consequences
Integrated XAI in TCS

Figure 0.1 XAI Design Framework Rationale

Aims and Objectives

The overall aim of this study is to develop a design framework for effective integration of XAI to improve user experience and HMT performance in TCS. The evolution of the research followed three stages, covered in 1 - 3 each with the following aims:

1. Develop a detailed understanding of the problems associated with integration of HMTs in expert systems and the consideration of the problem as a JCS.
2. Design and develop a framework for integrating XAI in TCS based on the 'Design Thinking' approach, covering the Initial development and initial proof of concept of a Design Thinking Framework for XAI in TCS.
3. Chapter 3 covers two main aims:
   a. Development and verification of performance assessment methods and effect relationship mapping for the XAI Design framework in the context of a real-world problem to support the development of design tools to guide engineers and designers in developing XAI.
   b. Evaluate the performance of the solution developed using the XAI framework in the context of a real-world problem. Conduct an experiment to understand the impact of explainable AI in TCS and provide evidence for the suitability of specific measures for task performance and quantify the relationships between task performance measures and constructs.

These aims were translated into research objectives and related research questions. These are summarized, along with specific tasks and hypotheses are outlined in Tables 0.1, 0.2 and 0.3 These tables provide an overview of the relationships between objectives, questions and tasks and further details, including measures and alternative hypotheses are provided in the relevant chapters of this document. To ensure that this dissertation is a coherent document, each chapter will contain a segue section to explain the context and importance of the particular aspects of the research, followed by self-contained research papers that address the specific aims of each chapter. In each of these segue sections, I include a real-world example of an AI drone co-pilot to demonstrate the context of the research to the reader and to provide insight into the importance and rationale of the study.

Table 0.1 Overview of Stage 1 Aim, Objectives, Research Questions and Tasks

Stage 1 – Understanding Time-Critical AI Systems

Aim: Develop a detailed understanding the problems associated with the integration of AI into human-machine interfaces in expert JCS, at the user interface level. (Stone, Ganapathy., 2021; Stone, Jessup, Ganapathy, Harel 2021).

Research Objectives	Research Questions	Research Approach/Tasks	Hypotheses	Reporting
1.1 Develop and validate a cognitive model of cardiac catheterization	**1.1.1.** How do experts make decisions in the context of human-machine teaming?	**1.1.a** Review existing cognitive models and develop a cognitive model of the human-machine JCS in the decision support system	N/A	Reported in this document
		1.1b Validate the cognitive model through application to specific scenarios and cognitive walkthrough.		Stone, Nelson, Ganapathy and Fendley (2021)
1.2. Apply AI to an Expert catheterization system.	**1.4.1** What are the practical issues surrounding implementation of AI in an expert system?	**1.4.a** Identify a specific element of the expert task with the potential for AI integration to enhance performance. **1.4.b.** Develop an AI model to assist human decision making.	AI can provide accurate advice in the context of a decision support system.	Stone & Ganapathy (2022) – submitted to *Smart Health*, October 2022)

Table 0.2 Overview of Stage 2 Aim, Objectives, Research Questions and Tasks

Stage 2 - A Design Thinking Framework for the Integration of Transparency Measures in Time Critical Decision Support.

Aim: Design and develop a framework for integrating XAI in TCS based on the 'Design Thinking' approach. Including the Initial development and proof of concept of a Design Thinking Framework for XAI in TCS.

Research Objective	Research Questions	Research Approach/Tasks	Hypotheses	Reporting
2.1 Develop a framework for integrating XAI in TCS based on the Design Thinking approach.	**2.1.1** What are the requirements for the assessment of XAI in TCS? **2.1.2** How can the key requirements for the design of XAI be mapped into the Design Thinking approach stages: Empathize, Define, Ideate, Prototype, Test **2.1.3** Can the Design Framework be used to successfully integrate transparency measures?	**2.1.a** Develop a taxonomy of design guidelines to facilitate designing and easy integration of XAI in TCS. **2.1.b** Develop a design framework that allows the requirements for XAI to be met and facilitates the integration to AI-based TCS. **2.1.c** Conduct Initial Heuristic Evaluation of the Design Framework on a candidate AI-assisted TCS.	N/A	Stone, Jessup, Ganapathy and Harel (2021)

Table 0.3 Overview of Stage 3 Aims. Objectives, Research Questions and Tasks

Stage 3 - Framework Implementation and XAI Assessment

Aim: Test the performance of the instantiated model developed by the Design Thinking Framework for XAI in the context of a real-world problem. Conduct an experiment to understand the impact of XAI in time-critical tasks, provide evidence for specific task performance measures, and quantify the relationships between task performance measures and constructs.

Research Objective	Research Questions	Research Approach	Hypotheses	Reporting
3.1. Develop and verify performance assessment methods and map effect relationships for the XAI Design Framework in the context of a real-world problem.	**3.1.1** How accurately and reliably can the impact of XAI on human trust in time-critical systems be measured? • Are subjective user assessment tools reliable methods to assess the effect of XAI in TCS? • Are eye-tracking assessment methods reliable for the assessment of the effect of XAI in TCS? **3.1.2** What are the effects of XAI information in TCS on: Task Performance, Trust, Workload and Situational Awareness? **3.1.3** What are the interactions between these effects?	**2B.1.a.** Identify direct and indirect measures for task performance and associated constructs. **2B.1.b.** Develop a generic assessment platform for a time-critical AI-assisted task **2B.1.c**. Conduct Human Subjects experiment to determine if physiological measurements i.e., eye tracking can predict and measure human trust in time-critical Human Machine systems	Subjective assessment methods can reliably be used to assess the effect of XAI in TCS. Eye-tracking assessment methods can reliably be used to assess the effect of XAI in TCS. XAI information affects task performance, trust, workload, and Situational Awareness in TCS and there are interactions between these.	Eye Tracking for XAI performance Assessment in Time-Critical Systems. (Stone and Ganapathy), Submitted to: *International Journal of Human-Computer Interaction* (November 2022)

3.2. Assess the performance of an implementation of the XAI design framework and determine the effect of XAI in time-critical systems.	**3.2.1** What is the performance of the XAI solution on task performance compared to the baseline solution? **3.2.2** What is the performance of the XAI solution on workload compared to the baseline solution? **3.2.3** What is the performance of the XAI solution on trust compared to the baseline solution? **3.2.4** What is the performance of the XAI solution on SA compared to the baseline solution?	**3.1.a.** Develop a representative assessment platform for a time-critical AI-assisted task **3.2.b**. Implement the XAI Design Thinking Framework on this task. **3.2.c.** Define required XAI Model conditions and Trust and Workload levels. **3.2.d.** Conduct a human-subjects experiment to assess the impact of transparency measures. **3.2.e** Map the relationships and dependencies between XAI and task performance, Trust, Workload, and SA.	It is expected that XAI information and its timing will affect task performance and the related measurement constructs, but this relationship is complex, and it is not known what the result of XAI information will be at this stage. Two-tailed testing will be used to determine statistical significance	Reported in this document

Chapter 1 – Understanding AI in Human-Machine Teams

In this chapter I focus on the exploration of the problem, covering the development of a deeper understanding of the requirements for human-machine teaming in an expert system, with the two specific objectives of developing and validating a cognitive model of cardiac catheterization, published in the Human Factors and Ergonomics in Manufacturing & Service Industries journal, 31(4), 360-374 (Stone & Ganapathy, 2022) and applying AI to a specific part of the catheterization problem submitted for publication in the international journal of Human-Computer Interaction.

The first of these problems I considered was the development of a cognitive model of cardiac catheterization. In the evolution of this research, it was determined that considering human-machine teams as a JCS can provide enhanced team performance, however this requires an understanding of both the task being conducted and the cognitive models employed. In addition to developing the cognitive model, it is important to understand how well it defines the cognitive processes it was designed to represent (Cen, Koedinger, & Junker, 2006). The original research validation study for the cognitive model of the cognitive model of cardiac surgery was initially reported in the paper 'Development of a Novel Hybrid Cognitive Model Validation Framework for Implementation Under COVID-19 Restrictions' in the journal *Human Factors and*

Ergonomics in Manufacture and Service Industries (Stone, Nelson, Ganapathy and Fendley, 2021). This study was conducted in the early phases of the COVID-19 pandemic of 2020 and, as such, the validation process had to rely on remote methods, as opposed to face-to-face interviews more traditionally used.

Taking the AI copilot example, this chapter relates to the steps required for the designer to develop a deep understanding of both the AI copilot and the user, i.e. the drone pilot. The first paper presented, establishes a cognitive model for a human operating in an expert system. The context presented here is not related to the drone co-pilot problem, however the time-critical and expert nature of the decision context are representative. This stage of the research relates to the early phases of XAI design where a deep understanding of the user, task, environment and the AI system requiring explanation is required. Including this element in the XAI design framework allows the designer to understand the cognitive processes of an expert user without needing to become an expert themselves.

Development of a Novel Hybrid Cognitive Model Validation Framework for Implementation Under COVID-19 Restrictions

Stone, P. B., Nelson, H. M., Fendley, M. E., & Ganapathy, S. (2021). Development of a novel hybrid cognitive model validation framework for implementation under COVID-19 restrictions. Human Factors & Ergonomics in Manufacturing & Service Industries, 31(4), 360–374. https://doi-org.ezproxy.libraries.wright.edu/10.1002/hfm.20904.

ABSTRACT

The purpose of this study was to develop a method for validation of cognitive models consistent with the remote working situation arising from COVID-19 restrictions in place in Spring 2020. We propose a framework for structuring validation tasks and applying a scoring system to determine initial model validity. We infer an objective validity level for cognitive models requiring no in-person observations, and minimal reliance on remote usability and observational studies. This approach has been derived from the necessity of the COVID-19 response, however we believe this approach can lower costs and reduce timelines to initial validation in post-Covid-19 studies, enabling faster progress in the development of cognitive engineering systems. A three-stage hybrid validation framework was developed based on existing validation methods and was adapted to enable compliance with the specific limitations derived from COVID-19 response restrictions. This validation method includes elements of argument-based validation combined with a cognitive walkthrough analysis, and reflexivity assessments. We conducted a case study of the proposed framework on a developmental cognitive model of cardiovascular surgery to demonstrate application of a real-world validation task. This framework can be easily and quickly implemented by a small research team and provides a structured validation method to increase confidence in assumptions as well as to provide evidence to support validity claims in the early stages of model development.

Introduction

The recent outbreak of COVID-19 (SARS-CoV-2) at the beginning of 2020, has caused 49.7 million individuals to become infected and caused 1.2 million fatalities globally as of November 2020 (World Health Organization [WHO], 2020). The extensive outbreak has prompted government intervention through widespread shutdown of non-essential businesses and services, implementation of social distancing guidance, and reallocation of resources and funds to better assist with viral mitigation and containment efforts (Ashraf, 2020). In addition to social and institutional shutdowns, economic downturn has also ensued due to loss of funding, lack of consumer spending, and uncertainty around the return to pre-COVID-19 normalcy. The ramifications have not only impacted the global economy but have also had a significant effect on the research community within public and private institutions (Ashraf, 2020).

The American Journal of Emergency Medicine has outlined specific areas for focus and guidelines for research during the COVID-19 pandemic (Haleem, Javaid, Vaishya & Deshmukh, 2020). Personal Protective Equipment (PPE) has been redistributed to those fighting the virus on the frontlines, participants and researchers have been prohibited from participating in in-person research (unless pertaining directly to COVID-19) and reassigned to staggered schedules or remote work to reduce the amount of face-to-face interaction and to maintain appropriate social distancing guidelines. Pertinent research in the healthcare field has also been largely suspended as an attempt to allocate the most physical and financial resources possible to fighting the virus and predicting future viral outcomes.

The face-to-face work of human factors researchers has been affected particularly hard by these restrictions, as this historically requires extensive human interaction to elicit information regarding cognitive and decision-making processes (Sy, O'Leary, Nagraj, El-Awaisi, O'Carrol & Xyrichis, 2020). Much of this work requires structured in-person studies to be conducted utilizing observational and probing techniques. Given that many current studies in Human Factors research are not related to battling COVID-19, they are not considered essential practices. In the case of healthcare, this problem is compounded as subject matter experts, especially those on the frontlines, are not readily available to participate in related studies, and non-essential research personnel are restricted from healthcare facilities, necessitating that test and evaluation procedures are moved online or remote (Sy et al., 2020).

During this time of rapid innovation around COVID-19, there is an opportunity to develop new and innovative tools and capabilities for remote human factors methods. Specifically, cognitive model development is an area of human factors research that is heavily reliant on face-to-face communication, both to develop models and perform validation and assessment studies. We believe there is a clear need for methods that allow us to develop and validate cognitive models with less reliance on in-person observations and face-to-face interviews and walkthroughs.

Cognitive models are widely used in psychology and cognitive engineering to understand where errors are made or how training systems can be developed to reinforce a user's cognitive model. They are a means for researchers to understand, describe and predict how individuals or teams perform cognitive tasks, such as information processing and decision making, and can highlight relevant cognitive states and actions with respect to a given task (Rupp & Leighton, 2016). Wagenaar, Reason and Hudson (1990) suggests that there are specific errors associated with specific cognitive states and that knowledge of the cognitive model can be used to reduce errors and improve human decision making.

Cognitive models aim to define how individuals or groups carry out tasks and detail the paths and states required to complete a goal. Hayes-Roth and Hayes-Roth (1979) develop a cognitive model to understand the nature of the planning activity from apparently rational to apparently chaotic decisions made when individuals develop plans. The model allows the researcher to abstract types of knowledge and decisions made in planning to parse top level thinking and enable simulation of the underlying process. Similarly, the Recognition-Primed Decision (RPD) model (Klein, 1993) examines the more abstract concept of naturalistic decision making. This model allows the researcher to relate outcomes of testing to the naturalistic decision model, which in turn can be used to confirm or reject an assertion about the type of decision-making being employed.

The application of cognitive models is addressed by Belkin (1984), who concludes that it is vital to understand the user's problem to build a cognitive model.

This explicit representation allows researchers to use computational techniques to quantify human cognitive performance through simulation (Cooper, Fox, Farringdon & Shallice, 1996). For cognitive modelling to deliver its full potential benefit to cognitive science, it is important to ensure sound methodological principles are used in the development. The key advantage of cognitive models is to create output that is repeatable and produces valid predictions. However, to create a consistent, valid prediction, it is important to validate the model.

There is a need for evidence in simulation and ergonomics science and studies of validity can provide this, giving additional credibility to the associated models (David, 2013; Stanton, 2016). We considered the questions asked by Landry, Malouin and Oral (1983) in determining our approach to validation: what does it mean for a model to be valid and, does validity refer to the output, structure, or modelling process? Validation can improve confidence in the methods used by human factors engineers and is an important step in the modeling of a system. (Stanton, 2016; Annett, 2002; Stanton & Young, 1999) As mentioned by Annett (2002), for ergonomic models it is essential to construct the validity of a model to ensure that the performance is consistent and predictive in nature. This provides credibility to the cognitive model for the given task. Also, it is important to note that cognitive models can be complex in nature and a validation method that works for one model, may not work for another (Strube, 2001).

Keehner, Gorin, Feng, and Katz (2017) discuss the general approach to validation of cognitive models, highlighting the requirement for an iterative, staged validation. The goal of any validation method should be to support claims and validity arguments about the specific models to which it is applied (Keehner et al., 2017). Similarly, Kane (2013) argues that validation is an ongoing and iterative process, and we believe a structured validation framework could have utility in the early stages of research where gathering validation resources is difficult or not cost effective. In this study we use the definition of validation as determining if the real system is represented by the model (Law & Kelton, 1991). This is achieved by mapping the system capabilities with the model representation. More robust, quantitative validation procedures may be required as models are developed, but here we seek to determine if the model meets the basic requirements of representing the underlying cognitive processes.

For this study, the safety-critical nature of cardiovascular surgery makes validation key to future implementation of this proposed model and heightens the need to find alternative means of cognitive model validation to enable progress under COVID-19 restrictions. This paper attempts to address the question - how can we validate a cognitive model in the age of COVID-19 and remote testing? We outline a streamlined validation process and explain how we adapted existing thinking while developing an understanding of the evolving COVID-19 situation and innovated new approaches to cognitive model validation.

Aims and Objectives

We aim to advance a simple, structured framework for cognitive model validation requiring no direct contact and minimal reliance on remote usability and observational studies. While this has been born out of the necessity of the COVID-19 response, we believe a hybrid validation framework for cognitive modeling can have broader application to support cognitive systems engineering. We believe this approach can lower costs and reduce timelines to initial validation and could allow identification of problems early in model development, potentially preventing problems further downstream. Our case study focuses on the validation of a cognitive model of cardiovascular surgery, given restrictions associated with COVID-19.

Scope

The intended output of this study is a Hybrid Cognitive Validation Framework to provide human factors researchers with a means to expedite initial validation with minimal resources. We propose a validation process based on analysis of existing literature, reflexivity cross check, and cognitive walkthrough within the research team. A case study implementation of the validation framework is presented to demonstrate application and provide example output from the framework. This analysis was contingent on the availability of a candidate cognitive model developed prior to the COVID-19 restrictions. Although these restrictions may impact the development of cognitive models, this study focuses exclusively on the validation of cognitive models in this context. The scoring system used in this study is only preliminary for framework confirmation through implementation of the framework and associated feedback on scoring usefulness and accuracy.

Development and Test Implementation of Validation Framework

In line with the aims and objectives of this study, we used the following process to develop and conduct a test implementation of the Hybrid Cognitive Validation Framework.

Formalization of Restrictions

COVID-19 restrictions are not uniform among countries, states and even cities and localities (Hale, Petherick, Phillips & Webster, 2020). As authorities balance the need for public safety with the desire to maintain economic activity, these regulations also vary over time. We therefore define specific conditions with which this validation framework is compatible. In addition to representing COVID-19 restrictions, these conditions also represent future situations under which the Hybrid Cognitive Validation Framework can be an effective tool for research teams. The restrictions used in this study are:

- The validation framework can be implemented by a research team consisting of a minimum of 2 individuals for performing reflexivity analysis requirements (Davies & Dodd, 2002), to collaborate and ensure checks and balances on the validation.
- No requirement for in-person intra-team meetings.
- No in-person contact with external Subject Matter Experts (SMEs) during the validation process.
- Communication with SMEs is limited to confirmatory questioning – no probing or enhanced analysis due to assumed lack of availability.
- The research team has access to validation resources, such as those available online.

These restrictions are in line with initial COVID-19 restrictions implemented by the state of Ohio during the initial phase of the COVID-19 response (DeWine, Husted, & Acton 2020).

Analysis of Existing Validation Techniques

American Psychological Association technical recommendations for psychometric tests defines the following four types of validity - Concurrent Validity, Predictive Validity, Construct Validity and Content Validity (American Psychological Association [APA], 1954). Predictive and Concurrent methods are considered together as criterion validity and refer to the ability of a test to accurately predict, either in advance or concurrently, a predetermined measure or characteristic (Cronbach & Meehl, 1955). By contrast, Construct Validity (APA, 1954) aims to determine if a test measures the underlying concept it aims to address (Middleton, 2019). Finally, there is Content Validity, which is established by demonstrating that test subjects are representative of the population of interest (Cronbach & Meehl, 1955). In this paper, we concentrate on construct and content validity, ensuring the model and its inputs are representative, rather than comparison of model outputs to known standards. These measures relate to the internal validity of the model and given the potential complexity of assessment, we do not focus on the external validity.

Prior to COVID-19 restrictions, the proposed validation procedure for the Cognitive Model of Cardiovascular Surgery, was to implement a method used by Craig, Klein, Griswold, Gaitonde, McGill, and Halldorsson (2012) to validate a cognitive model of laparoscopic surgery. This method includes three validation stages, including data collection through SME interviews, construct encoding and comparison, and reflexivity phases including reassessment by additional participants who did not take part in the initial assessment. This is a modified version of an evidence collection and model validation developed in the knowledge audit method (Militello & Hutton, 1998). The process required multiple data elicitation procedures and cross comparison with subject matter experts with explicit procedures aimed at minimizing researcher bias.

The restrictions adopted in this study exclude face-to-face, concurrently gathered validity measures, and drive us toward remote, asynchronous measures, which, although easier to collect are somewhat harder to infer target cognitive processes from (Embretson, 1983). We considered several approaches to establish construct and content validity for the HCOG framework. Thoromon, Salmon & Goode (2019) use interview data as a reference standard to evaluate the validity of a near-miss reporting form. In this study, we adopt a similar approach, utilizing the cognitive walkthrough as our interview reference standard for empirical validity assessment of the HCOG model. Silva, Vieira, Campos, Couto, and Ribeiro (2020) validate a Descriptive Cognitive Model for predicting performance in Low-Code Development platforms. In this study, the model validation is achieved by comparison of knowledge-based and systems-based descriptions and by analysis of the model against specific tasks, appropriate to the Low-Code Development Platform. Stanton and Barber (2005) validate the Task Analysis for Error Identification, demonstrating improved performance compared with Heuristic evaluation, an approach developed by Stanton and Stevenage

(1998). This method showed good reliability and concurrent validity for the Task Analysis for Error Identification technique. Cornelissen, McClure, Salmon, and Stanton (2014) consider the validation of a formative method, concentrating on cognitive work analysis, noting the importance of pooling results of multiple analysts in establishing validity. This reinforces the importance of the requirement for multiple researchers to conduct our validation framework.

Vinod, Tang, Oishi, Sycara, Lebiere and Lewis (2016) utilize Markovian modelling to represent humans and build task simulations to compare outcomes with potential human action. This was initially considered a strong option to provide objective, quantitative basis for evidence generation in the Hybrid Cognitive Validation Framework, however the complexities of this approach and expert nature of surgeons meant this approach was discounted.

The argument-based approach to validation (Kane, 2013) aims to minimize complexity in the validation process while still evaluating claims and providing evidence to support them. The argument-based approach is based on early construct validity models (Cronbach and Meehl, 1955) which details three general principles for validity:

- The focus of the validation is on the interpretation of the output rather than the output itself.
- Validation is part of an ongoing research program.
- The proposed interpretation of the output is subject to critical evaluation.

The argument-based validation framework (Kane, 2006, 2013) also requires two argument types to support validity: An interpretive/use argument (IUA) and a validity argument. The IUA argument specifies the claims that are to be evaluated in the validation, while the validity argument is used to evaluate the interpretation of validation scoring. The argument-based method claims if these arguments are clear, coherent, and complete, and the inferences reasonable and assumptions plausible, a model can be said to be valid. We will use these definitions as the basis for the argument-based elements of our validation framework.

The cognitive walkthrough is a structured review method for conducting usability assessments early in the design cycle of a product (Lewis, Polson, Wharton, & Rieman, 1990). It involves the generation of task scenarios and explicit assumptions regarding the user population and context of use. This method was initially developed to assess the usability design performance of user interfaces, but we believe it can be adapted to establish construct validity of models . The method requires definition of the user along with sample tasks or scenarios and action sequences that are compared with an implementation of the user interface (Lewis et al., 1990). In adapting this method to the validation of cognitive models, we establish sample tasks and incorporate these into credible vignettes or scenarios based on the implementation associated with the cognitive model. We ask the reviewer to determine if the cognitive paths and states in the model are representative of the decisions associated with the tasks in example scenarios, and to walkthrough the scenario, task-by-task and compare to the cognitive model of cardiovascular surgery.

The intended use of the model of cardiovascular surgery is to predict the surgeon’s workload through indirect assessment of cognitive state, linking the paths taken through the model to periods in a procedure where workload is high or low. Ideally, a method to evaluate concurrent validity of the prediction made by the model would be implemented, but given the restrictions due to COVID-19, this is not possible. We are aiming to demonstrate construct validity and content validity. Specifically, we aim to establish the ability of the model to represent the underlying concept (construct validity) and the representativeness of the model to the target population, in this case cardiovascular surgeons. For the purposes of this study, we utilize existing Cognitive Models in the decision ladder and recognition-primed decision model. These models have been widely used and are subject to validation (Rasmussen, 1979; Lintern, 2010; Soh, 2007.) Rather than focus on internal validation of the model structure, this study focuses on developing a validation approach to answer the question ‘Is a specific model representative of the cognitive tasks for which it is built?’. We therefore propose a concept for a hybrid cognitive walkthrough (Lewis et al., 1990), argument-based validation method (Kane, 2006) to establish construct validity, augmented with reflexivity analysis (Davies & Dodd, 2002) to establish content validity. We believe including both the walkthrough analysis and argument-based methods provides a broad but flexible and adaptable basis for the validation of cognitive models that integrates into early iterations of model development and can be used to derive requirements for more complex assessments as well as providing evidence for model validity.

Define Hybrid Cognitive Validation Framework

This concept was developed into a specific validation framework detailing specific tasks to establish both construct and content validity. The tasks are representative of those in the donor validation methods, highlighting surrogate tasks where the restrictions of this study limited the scope of initial application. The expected outputs and interpretations are also defined.

Test Implementation of the Validation Framework

The resultant framework was implemented in a case-study validation of a cognitive model of cardiovascular surgery. The wider context of the development of this cognitive model is the development of a Decision Support System (DSS) to improve performance and reduce risk in cardiovascular surgery. Woods (1985) proposes that integrating machine and human cognitive systems is the key to the application of such a Decision Support System (DSS) and that considering DSS as a unified Joint Cognitive System (JCS), as opposed to a stand-alone machine-based system, enables enhanced understanding of the goals and responsibilities of the system.

Hybrid Cognitive Validation Framework

The Hybrid Cognitive Validation Framework consists of three core tasks, applied using an associated use case to illustrate the context for wider implementation. The three core tasks are: 1) Define objective and Interpretation framework (argument-based method), 2) Walkthrough analysis 3) Reflexivity analysis.

Task 1 is closely aligned to the two-stage argument-based validation method (Kane, 2013). The IUA task (Task 1A), requires the definition of a specific validation question and a definition of the purpose and scope of the validation framework implementation. The Validity task (Task 1B) requires the definition of specific interpretations of model validity assessment. These tasks should be conducted prior to implementing the remaining tasks within the validation framework implementation to prevent bias or fitting interpretations to align with results of the later analyses. Task 2 is the cognitive walkthrough validation method. This is aimed at confirming the cognitive paths and states within the model align with specific, credible cardiovascular surgery scenarios, by comparing defined tasks with the cognitive paths and states in the model. This method was developed to enable implementation without the need for face-to-face contact. This lack of direct exposure of SMEs to a cognitive model walkthrough is a key limitation of the Hybrid Cognitive Validation framework, however our walkthrough analysis approach consists of two stages. Firstly, Task 2A, scenario development, where detailed vignettes of representative situations are generated. Multiple scenarios should be developed to provide greater variation of cognitive states and pathways to ensure robust validation of the cognitive model. Secondly, a walkthrough stage where Subject-Matter Experts utilize these vignettes, with reference to the cognitive model to determine how representative the states and pathways compare to the vignette requirement. Rather than require Subject Matter Experts to record outcomes at each decision point and attempting to encode potentially incomplete or inaccurate data, we propose a four-level qualitative scoring system to enable a simplified assessment. Task 3 is an implementation of the Reflexivity analysis (Davies & Dodd, 2002). This stage aims to ensure model validation procedures and assessments are supported by suitable evidence and that SMEs used are qualified and documented. Clarity, coherence, and completeness are key to all tasks in this framework, in line with requirements defined by Kane (2013). The resultant Hybrid Cognitive Validation Framework is detailed in Table 1A.1.

Table 1A.1 – Hybrid Cognitive Validation Framework

ID	**Sub Task** (Including responsibility) Task definition	**Qualitative Validation Criteria** Individual score(0-3pts) Total task score (6 pts) Framework score (/18pts)	**Objectivity (Reflexivity)** Individual Score (0-2pts) Total task score (4 pts) Framework score (/12pts)	**Overall Score /Evidence** Task Score (/10) Framework (/30)
Task 1 Define the Validation Assessment Objective and Interpretation Framework (Argument-Based Method)				
T1A	**IUA Task** (Research Team) Define the validation assessment question and purpose including the purpose and scope of the validation.	The validation task is clearly stated and in line with the limitations of the framework. **(+3 if true)**	Validation objective clarity and coherence with interpretations assessed by second researcher. **(+2 if true)**	Written confirmation of assessment. **(Total /5)**
T1B	**Validity Argument Task** (Research Team) Define validation scoring interpretations framework: Establish the implications and assumptions of the validations.	Inferences are reasonable, assumptions are plausible and underpinning descriptions and evidence are clear, coherent, and complete. **(+3 if true)**	Validation objective clarity and coherence with interpretations assessed by second researcher. **(+2 if true)**	Written confirmation of assessment. **(Total /5)**
TASK 2 – Walkthrough Analysis Generate Scenarios, test matrix. Conduct cognitive walkthrough.				
T2A	**Scenario Development Task** (Research Team) Generate scenarios that are representative of model use cases to enable cognitive walkthrough.	**Scenario** Assessment (Subject Matter Expert) Scenarios are representative of operational situations (routine, emergency, complications). **(+1 if true)** Classes of patient and procedure types are representative and appropriate (age, underlying condition, gender). **(+1 if true)** Scenarios clearly represent the complete cognitive tasks identified in the cognitive task matrix **(+1 if true)**	Scenarios confirmed by second researcher. **(+1 if true)** Scenarios approved by external subject matter expert. **(+1 if true)**	Detail scenarios used in the validation. Provide Cognitive Task Matrix. **(Total /5)**

ID	**Sub Task** (Including responsibility) Task definition	**Qualitative Validation Criteria** Individual score(0-3pts) Total task score (6 pts) Framework score (/18pts)	**Objectivity (Reflexivity)** Individual Score (0-2pts) Total task score (4 pts) Framework score (/12pts)	**Overall Score /Evidence** Task Score (/10) Framework (/30)
T2B	**Asynchronous Cognitive Walkthrough** (Subject Matter Expert) SME to Conduct cognitive walkthrough. if the model is applicable to specific tasks and infer generalizability to tasks not identified explicitly in the model development.	**Cognitive Model Assessment** There are significant gaps in the representation cognitive states and pathways for many tasks. **(+0 if true).** The model meets the basic requirements of representing the underlying cognitive processes but may have some gaps **(+1 if true)** Cognitive states, actions and pathways were clear, complete, and coherent and representative of all surgical phases in one scenario. **(+2 if true)** Cognitive states, actions and pathways were clear, complete, and coherent and representative in at least two scenarios. **(+3 if true)**	Assessment of model representativeness and cognitive walkthrough confirmed and assessed by second researcher. **(+2 if true)**	Record walkthrough results and detail cognitive states actions and pathways as appropriate. **(Total /5)**
TASK 3 - Reflexivity Assessment: Determine if appropriate data was collected to support model development?				

T3A	**Cross-check qualifications**
	(Research Team)
	Were appropriately qualified subjects used in the model development?
		Subject(s) has/have no experience or expertise in the field.**(0 pts)**
		Subject(s) has/have some training and minimal experience – less than 2 years. **(1 pts)**
		Subject(s) is/are experienced practitioner(s) and currently practicing – 2-5 years.**(2 pts)**
		Subject(s) is/are considered expert(s) and currently practicing – more than 5 years.**(3 pts)**
			Assessment of subject qualifications performed by second investigator/or qualifications and experience is known. **(+2 if true)**	Qualifications and experience of those contributing data is recorded.
				(Total /5)

T3B	Was the data collection process robust and complete during model development?	Data collection covered the breadth and depth of the model and a representation can be made. **(+1 if true)** Data collection conditions were consistent across all participants.	The debrief was conducted by a second researcher, not present in the data collection. **(+1 if true)** The debrief was conducted within one	Identify any data collection techniques used along with records of data collection. **(Total /5)**

ID	**Sub Task** (Including responsibility) Task definition	**Qualitative Validation Criteria** Individual score(0-3pts) Total task score (6 pts) Framework score (/18pts)	**Objectivity (Reflexivity)** Individual Score (0-2pts) Total task score (4 pts) Framework score (/12pts)	**Overall Score /Evidence** Task Score (/10) Framework (/30)
		(+1 if true) Relevant data collection was used without omission or bias. **(+1 if true)**	week of the data collection. **(+1 if true)**	

The reviewers identified to conduct the cognitive walkthrough are provided with the following instructions on the implementation of the method:

- This is a walkthrough validation of the states and paths detailed in the cognitive model of cardiovascular surgery.
- The aim is to establish validity through comparison with tasks and decision points in the scenario.
- Familiarize yourself with the Cognitive Model under review (Figure 1)
- Familiarize yourself with the Hybrid Cognitive model validation Framework (Table 1)
- Conduct a task-by task walkthrough of the validation scenario (Table 4 and 5) with reference to the cognitive task matrix (Table 3).
- For each task identified in the walkthrough scenario, record the paths taken through the cognitive model and identify where decisions and cognitive states and paths do not match.

These instructions are designed to be sent electronically and can be followed up with a discussion with the reviewer to clarify any elements of the walkthrough task. The research team can then utilize the results of the walkthrough analysis, in conjunction with the validation interpretation framework (Table 1A.2) to assign validity scores to the model. This is an initial implementation of the walkthrough analysis method to support cognitive model validation and it is expected that the guidance for reviewers will be expanded based on the feedback received during this study.

Validation Framework Implementation and Scoring

The Hybrid Cognitive Validation Framework (Table 1) should be used as part of an iterative, scalable validation process with tasks completed sequentially. The model development stage is not scored but is included to indicate the chronology of the development within the validation process.

The argument-based analysis, walkthrough analysis, and reflexivity analysis tasks, detailed in Table 1, each have two sub-tasks with a potential for 5 points. The success criteria, or Validation (V) element scores (0-3 points) are summed with the objectivity (O) scores (0-2 points) to give an overall Validation Framework (VF) score (0-5 pts). This gives a potential score of 5 points for each of the six tasks and sub-tasks and a total of 30 points for each implementation of the validation framework, with 18 points attributable to validation success criteria and 12 points to objectivity scoring. The inferences derived from this scoring are variable, dependent on the validation task, defined in the argument-based validation task (Kane, 2013). For a model to be considered 'Good', some objectivity analysis should be undertaken. For this reason, the threshold for 'good' should be above 18 out of 30, hence even if the model achieves a score of 18/18 on the 'validation criteria element, the interpretation threshold should require a score above this to ensure that some objectivity analysis is completed.

Cognitive Model for Case Study

A generic cognitive model of cardiovascular surgery was developed with cardiovascular surgeons from the Miami Valley Hospital, Dayton, OH using a Cognitive Task Analysis method (Craig et al., 2012), along with unstructured interviews and cognitive walkthrough. Rather than develop a new model from scratch, existing cognitive models were considered to represent distinct phases in surgery with the aim of developing a more robust model with existing evidence. This was augmented with an analysis of existing procedural documentation, specifically the COCATS 4, Task Force 10 training procedure (King, Babb, Bates, Crawford, Dangas, Voeltz, & White, 2015) and a stent fitting procedure (Stent: Purpose, Procedure, and Risks, 2017). The research team observed cardiovascular procedures at Miami Valley Hospital, conducting pre-op and post-op interviews with the surgeon to define cognitive states and actions. Finally, procedures carried out on the low-cost cardiac catheterization simulator, were conducted by the research team to understand first-hand the complexities of the cognitive task.

A unified model representing naturalistic, analytical, and mixed decision types was synthesized by modifying and combining existing cognitive models. The RPD Model (Klein, 1993; 1999) forms the representation of expert responses to routine situations with representation of analytical decisions in non-routine scenarios based on the Decision Ladder (Rasmussen, 1974). There are 'shortcut' paths in the model representing mixed decisions, aligned with heuristics of experienced surgeons. This dynamic cognitive model represents the relative distributions of normative and descriptive modeling as determined by Craig et al. (2012).

The RPD model (Klein, 1993; 1999) is used to understand how people make quick decisions in complex situations, particularly in expert domains. The model is derived from research into intuitive decision making and assumes use of prior knowledge and pattern recognition to make decisions. There are two key elements in the RPD model: firstly, the way a decision maker assesses a situation and recognizes a suitable course of action, and secondly how a course of action is imagined, and potential outcomes evaluated. Both of these elements are dependent on the ability to recognize both features of the situation and corresponding actions (Klein, 1993). This model is more typical of the advanced or expert decision maker, as higher situational awareness, and ability to predict outcomes based on experience enables this type of nuanced, heuristic decision making (Klein, 1993). The decision ladder model (Rasmussen, 1974) is representative of both the analytical decision-making paradigm and the heuristic, intuitive paradigm. In this model, the rational decision process follows the outer path of the 'ladder' whereas the heuristic decision process may start and finish anywhere in the model appearing as 'shortcuts' (Rasmussen, 1974). These models were combined with a single start point as in the Complex RPD Strategy Model (Klein, 1993; 1999). The three cognitive paths through the model are labelled P1 (Intuitive Decision Making), P2 (Mixed Decision Making), and P3 (Analytical Decision Making). Cognitive states are denoted by rounded boxes and actions in the square cornered boxes. The decision node represents where a surgeon's cognitive state can switch from intuitive to analytical. An additional external decision support node (labelled D2, Figure 1A.1) was introduced in the analytical decision phase to illustrate the potential for the surgeon to increase interaction with the other members of the team to determine a course of action in a complex, unfamiliar situation. The Cognitive Model of Cardiovascular Surgery (Stone, 2020) is shown in Figure 1A.1.

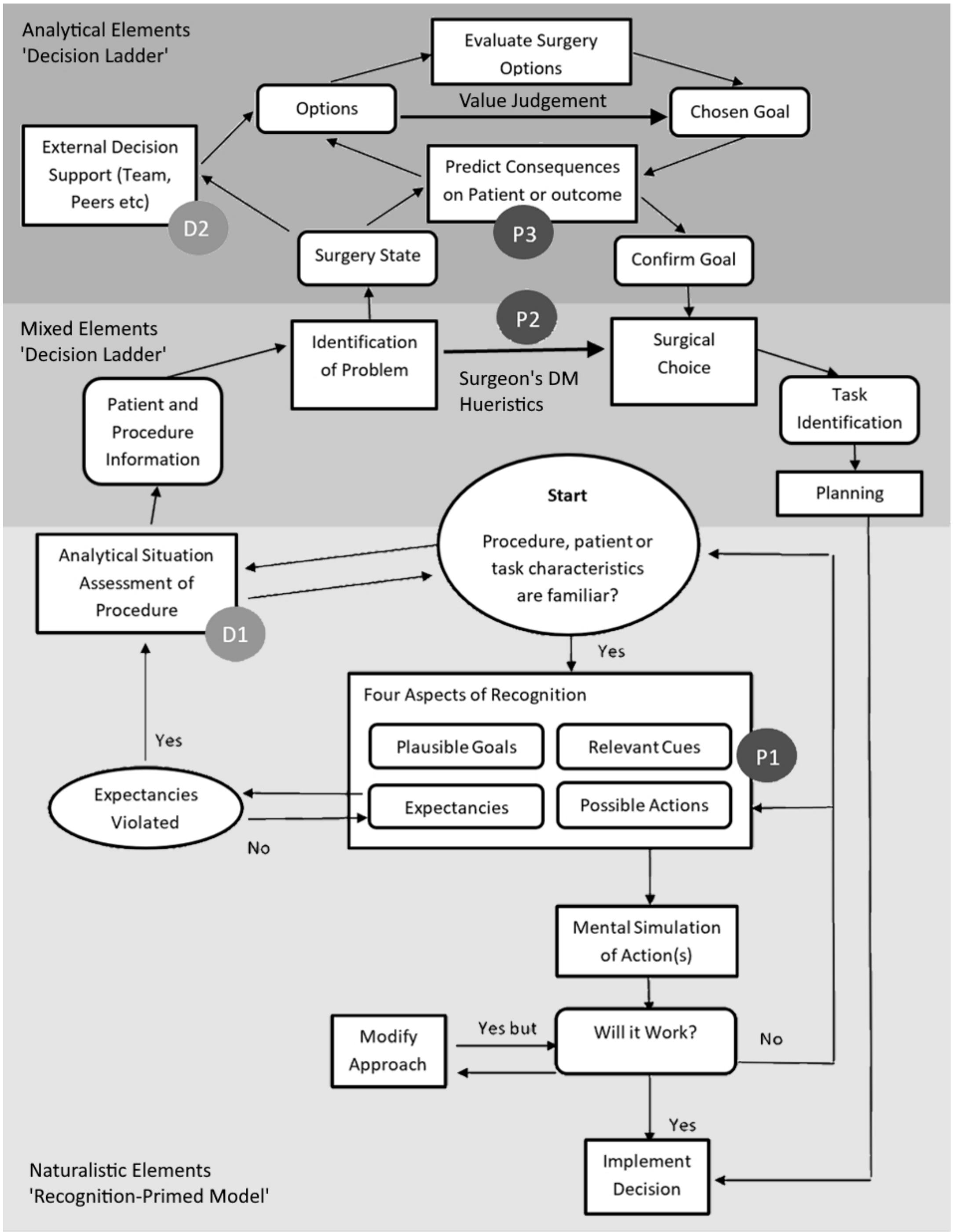


Figure 1A.1 – Cognitive Model of Cardiovascular Surgery

Results

The following case study implements the three tasks defined in the Hybrid Cognitive Validation Framework (see Table 1A.1). These tasks are implemented in the validation of a cognitive model. The scores for each task are detailed in Tables 1A.6 and 1A.7.

Task 1 – Argument-Based Validation

Task 1A - Define Validation Assessment Question and purpose (Argument Based)

Our Validation Question is "Does the Cognitive Model of Cardiovascular surgery represent the underlying cognitive processes of the cardiovascular surgeon"? This case study focusses on the initial validation of a developmental cognitive model, examining the approaches, assumptions and techniques that contribute to the model as well as an asynchronous, scenario-based walkthrough assessment of the model representation and appropriateness.

Task 1B - Define Validation Interpretations Framework (Argument-Based Validity)

We defined a five-level interpretation hierarchy with inferences attributable to both the validity and reflexivity scores. The validation score interpretation framework is shown in Table 1A.2.

Table 1A.2 – Validation Score Interpretation Framework.

Validation Status	Interpretation	Criteria
High Overall Validity,	The model can be said to have a high validity for an early developmental model and is suitable for implementation in initial research studies only. Model validation should continue as part of an ongoing, iterative design process. The model has high validity scores across all validation tasks. There is good evidence that the underlying assumptions are valid, data collection techniques are sound and researcher bias has been addressed through reflexivity assessment. The model has been demonstrated to be representative. The inferences of the Interpretation framework are reasonable and the assumptions plausible and the definitions of are clear, coherent, and complete.	Total VF[1] Score ≥ 25 (max = 30) Interpretation Score ≥ 7 Representation Score ≥ 7 Reflexivity Score ≥ 7 Min Success Criteria Score ≥ 2 Min objectivity Criteria Score ≥ 1
Good Validity, **Poor Reflexivity** **Well defined Interpretation**	There is evidence that this model has good validity for an early developmental model and clear, complete, and coherent interpretations were defined. This model may be useful for implementation in initial research studies, however reflexivity scores were low so there is a remaining caveat on the potential for researcher bias. Further external confirmation of the model is required to be used with confidence.	Total VF Score ≥ 19 (max = 30) Interpretation Score ≥ 6 (60%) Representation Score ≥ 6 (60%) Reflexivity Score ≤ 6 (60%) Objectivity score ≤ 6 (60%)
Good Validity, **Good Reflexivity Poorly defined Interpretation**	There is evidence that this model has good validity for an early developmental model and reflexivity assessments have been complete. This model may be useful for implementation in initial research studies, however interpretation frameworks were not provided so findings should be treated as somewhat speculative until an interpretation framework is defined.	Total VF Score ≥ 19 (max = 30) Interpretation Score ≤ 6 (60%) Representation Score ≥ 6 (60%) Reflexivity Score ≥ 6 (60%)
Poor Validity, **Good Reflexivity Well defined Interpretation**	Underlying data and assumptions used in the development of this model were somewhat unclear, incomplete, or incoherent. Researcher bias has been assessed and interpretations are clear, however caution should be used when implementing this model outside research team development activities.	Total VF Score ≥ 19 (max = 30) Interpretation Score ≥ 6 (60%) Representation Score ≤ 6 (60%) Reflexivity Score ≥ 6 (60%)
Poor Overall Validity	There are significant problems with two or more of the validation criteria. Further validation effort is required before the model can be used, even in initial research investigations.	Total VF Score ≤ 19 (max = 30)

[1] Validation Framework (VF) score defined on page 11 of the main text.

Task2 – Representation Assessment (Cognitive Walkthrough Analysis)

Task 2A - Generate scenarios

Scenarios representative of the use case were developed to enable cognitive walkthrough. To ensure scenarios generated for the walkthrough were representative of different cognitive states of surgeons throughout the process, a matrix of surgical cognitive tasks corresponding to Intuitive, Mixed and Analytical Decisions was defined, see Table 1A.3. This matrix was used to define each stage of the scenario. The scenario used in this case study is outlined in Table 1A.4.

Table 1A.3 Cognitive Task Matrix for Generic Cardiovascular Surgery Scenarios

Cardiovascular Surgery Subtasks and associated hierarchical cognitive decision examples

Sub-tasks	Intuitive example	Mixed Example	Analytical example
Procedure Planning	The combination of patient characteristics and procedure type and complexity are familiar.	The patient and procedure type are largely familiar, but some anomalies discovered but within situations governed by analytical heuristics.	The patient characteristics and/or procedure type are unfamiliar and additional cognitive resource are assigned to develop an analytical solution.
Patient Preparation	Patient responds to sedative and initial preparation as expected.	Some anomalous response but within experiential reference and responds as expected.	Patient response is not expected, and initial remedial measures are unsuccessful.
Catheter selection	Patient size, condition and vascular geometry are familiar and correspond to experience of successful catheterization	Some patient characteristics are unfamiliar but generally within expectations and can be extrapolated from experience.	Complex vascular geometry, narrowing or blockage requiring reassessment of entry or catheter choice (size shape etc.)
Catheter Insertion	Vascular location and orientation are predictable and catheter insertion at chosen site is successful.	Some difficulty in locating appropriate insertion point but alternatives discovered and implemented.	Problems inserting catheter due to depth of vascular network. Potential bleeding occurs or site reassessment required.
Catheter maneuver	Catheter behaves as expected and catheterization task can be completed	Some difficulty in maneuvering the catheter to the desired site but behavior is predictable and corrected easily	Catheter does not behave as expected and anomalies cannot be understood or corrected with existing heuristics.

Table 1A.4 – Case Study Validation Scenario Definition

Scenario phase	Description
Procedure Type:	Planned, routine procedure – Stent fitting to correct narrowing of coronary artery.
Patient Definition:	58-year-old male, 6 feet tall, with a body mass index of 25
Patient Preparation Complications:	none
Catheter Selection Complications:	The Fluoroscopic imaging of the patient reveals a narrower than expected vascular structure for a man this size and the initial expectations on catheter size and shape are violated.
Catheter Insertion Complications:	none
Catheter Maneuver Complications:	The catheter maneuver is unsuccessful due to the narrowed vascular geometry.

Task 2B - Cognitive Walkthrough

A cognitive walkthrough of each scenario/vignette was conducted with reference to the cognitive model. The cognitive model was scored for representativeness against the cognitive paths, actions, and decision points identified in the model in Figure 1A.1. The output from the cognitive walkthrough is detailed in Table A1.5 and both initial and reflexivity scores are provided in Tables 1A.6 and 1A.7.

Table 1A.5 Case Study Validation Scenario Walkthrough with Result (in bold)

Walkthrough phase	Description
Step 1	A fifty-eight-year-old male has been diagnosed with a narrowing of a coronary artery and has been scheduled for a stent fitting to correct the condition. The patient characteristics are within the experience of the cardiovascular surgeon and planning elements are routine. **This follows path P1 in the cognitive model and is representative of Intuitive decision making.**
Step 2	The preparation is conducted by the anesthetist and the patient responds as expected. **This follows path P1.**
Step 3	This vascular geometry is assessed creating a decision point at D1. The vascular geometry is found to be narrower than expected in line with the scenario definition, resulting in the surgeon using their experience and associated heuristics to reselect an appropriate catheter based on this new information. **This is a mixed Intuitive-analytical decision, following path P2.**
Step 4	A suitable insertion point is easily found but catheter insertion task is somewhat harder than expected. The surgeon intuitively corrects and quickly achieves a successful insertion without requiring consultation with other team members. **This is within the bounds of the Recognition-Primed Decision model – not requiring analytical heuristics so still follows path P1.**
Step 5	The maneuver of the catheter to the procedure site is unsuccessful in line with the scenario definition. The surgeon corrects position and tries again but is still unsuccessful. **Expectations are violated and path P3 is adopted. The surgeon may need to consult with external decision support and generate multiple options such as a different catheter or entry point. These options are evaluated against through a value judgement in the analytical level of the model.**

Task 3 Reflexivity Assessment

Task 3A - Were appropriately qualified participants used in the model development?

Underpinning assumptions for the model were developed in consultation with a cardiovascular surgeon with over five years' experience of cardiac surgery. Assumptions and representations were also generated through procedural observation and consultation with procedures and tasks outlined in COCATS 4, Task Force 10 training procedure (King et al., 2015).

Task 3B - Was the data collection process robust and complete during model development?

The rationale for excluding computational validation methods was explained in the development process. The breadth and depth of data sources were highlighted in the model development. The model development utilized existing research to establish a credible solution. The development steps are explicitly linked to data collection activities and establish a clear rationale for model development.

Score Summary Results and Interpretation guidance

The preliminary scores derived from the initial Hybrid Cognitive Validation Framework assessment of the case study model are summarized in Table 1A.6. It can be seen from these results that there is a '0' objectivity score as the assessment was conducted by the primary researcher. In this case, the Interpretation guidance would be that this model has poor validity, despite high validation scores, as no objectivity analysis was complete at this point. The rationale for this decision is outlined in the Section 'Validation Framework Implementation Scoring'. The validation procedure was repeated by a second researcher to show the improvement in scoring associated with the objectivity scoring element and the results are given in Table 1A.7.

Table 1A.6 – Hybrid Cognitive Validation Framework Score Summary – Primary Researcher

	Success Criteria score	Objectivity Score	Total Score
Task 1 - Argument-Based Analysis	6/6	0/4	6/10
Task 2 - Walkthrough Analysis	5/6	0/4	5/10
Task 3 -Reflexivity Analysis	6/6	0/4	6/10
Total (Vertical Sum)	17/18	0/12	17/30

The preliminary validation framework scores from all three validation tasks and associated sub tasks are collated and summed to use in conjunction with the Interpretation Framework, Table 1A.2.

Table 1A.7 – Hybrid Cognitive Validation Framework Score Summary including Objectivity Rating

	Success Criteria score	Objectivity Score	Total Score
Task 1 - Argument-Based Analysis	6/6	4/4	10/10
Task 2 - Walkthrough Analysis	4/6	4/4	8/10
Task 3 -Reflexivity Analysis	5/6	3/4	8/10
Total (Vertical Sum)	15/18	11/12	26/30

The complete validation framework scores from all three validation tasks and associated sub tasks are collated and summed to use in conjunction with the Interpretation Framework, Table 1A.2.

Second researcher variance in the implementation of the validation framework scoring stemmed from tasks 3B and 4B in both Success Criteria and Objectivity components. Task 3B demonstrated the basic requirements of the framework were met, but only on a single scenario. Additionally, this was not found to be completely representative of all surgical phases. Task 4B alternatively, demonstrated that data collection covered the breadth and depth of the model for a representation to be made and collection conditions to be consistent, however there was insufficient evidence to determine data was collected without bias or omission. The second researcher was not present for a debrief within one week of data collection, limiting the Objectivity score to a score of 1.

Overall, the second researcher scoring lowered the success criteria scoring from 17 to 15 but improved the objectivity scoring from 0 to 11. This underlines the importance of second researcher involvement in the assessment phase. The VF score resulting from the second researcher implementation of the case study was 26. When applied to the interpretation guide (Table 1.2), the cognitive model used in this case study was determined to have ‘High Overall Validity’.

Discussion and Conclusion

The Hybrid Cognitive Validation Framework developed in this study was demonstrated in a case study under COVID-19 restrictions. This framework was able to establish construct validity and content validity in line with expectations given input of a second researcher to evaluate initial validation assessments. This framework provides a structured means to approach initial validation studies where traditional validation resources are either restricted or projects do not have the means to access them. Including this simple validation process in projects can help to determine early on if models have potential validity and help to develop model inputs and requirements more accurately.

The Hybrid Cognitive Validation Framework was able to provide evidence-based validity scores and associated interpretations given the COVID-19 restrictions with no requirement for face-to-face validation assessments. This study demonstrated that the framework can be easily implemented by a small team with limited resources, highlighting the importance of objectivity elements in the method. The argument-based validation elements used are comparable with those outlined by Kane (2013). The reflexivity assessments are comparable with the process outlined by Davies and Dodd

(2002), however these are contingent on review by a second researcher or external subject matter expert.

A key limitation of this framework is the reliance on retrospective, scenario-based walkthroughs to replace concurrent assessment during procedures. The impact of this limitation is hard to quantify but while it may be considered to have lower credibility compared to existing methods, it fulfills the goal of model validation requiring no direct contact and minimal reliance on remote usability and observational studies. In this study, we discuss the validation of cognitive models under COVID-19 restrictions. These restrictions will also impact cognitive model development; however, this was not addressed in this study as our model was developed before the restrictions.

This initial case study implementation of the Hybrid Cognitive Validation Framework has shown promise to enable early validation with limited resources, the development process was rapid, and many lessons were learned along the way. Adapting to new ways of working has been necessitated by the COVID-19 restrictions that have become the new normal. The framework detailed in this study should be considered the starting point to refine and adjust as new evidence and experience informs its development, particularly regarding tuning of the validation scoring elements and their links to the defined implications.

The Hybrid Cognitive Validation Framework allows researchers to rapidly establish initial validity for cognitive models, especially given the restrictions associated with COVID-19, and other situations where face-to-face contact with SMEs may be limited, due to location or availability of the experts. This implementation focuses on the initial validation of the cognitive model of cardiovascular surgery, but we believe this approach could be easily used to validate any cognitive model, particularly those relating to decision making in complex environments, where access to expert input is limited, such as military, petrochemical, medical, or aviation. The keys to broader implementation of this method are the ability to establish the interpretive/use argument and validity arguments defined in the argument-based validation method along with the development of credible scenarios with tasks that represent the potential cognitive states and paths defined in the cognitive model (Kane, 2006; 2013).

While we believe this validation framework has utility under the circumstances identified, it is more prescriptive than other methods discussed in this paper and does not cover criterion validity. Ideally, we would like to establish concurrent validity through correlation tests between the predicted model state and concurrent assessments of a surgeon's employed mental model assessed by SMEs. This approach would be potentially less subjective and easier to compare using quantitative tests. While this establishes further validity evidence, it is potentially much more complex and requires access to resources that are not compatible with the rationale of an early-stage, low-cost validation approach presented here. The use of interview data as a reference for model validity (Thoromon, Salmon & Goode, 2019) would provide a potentially more robust means to gather validity data. A simulation approach, as employed by Vinod et al. (2016) could potentially establish more objective validity data but has potential corresponding validity issues arising from the simulation of human agents in a specialized role. Stanton (2016) notes that small assessment groups are frequently a problem with validation in Human Factors Engineering, and this case study was no exception, limiting the confidence in the conclusions until further research can be completed.

Due to time constraints, this study only considers a single scenario for the walkthrough analysis validation with a single SME. The next stage in the development of this validation framework will be to develop a more extensive set of scenarios for the cognitive walkthrough to establish more evidence for model validity. To enable this, it is expected that a more detailed scoring framework will be required to bridge the gap between the reviewer's cognitive walkthrough responses and the validity scores assigned in the validation interpretation framework.

COVID-19-like restrictions also have the potential to disrupt the development of cognitive models, prior to, or in parallel with validation activities. Future studies should address potential methods to address the impact on model development.

The detail in the instructions given is another potential limitation of this study, further development of the instructions and the presentation of the cognitive walkthrough task is important to ensure clarity and consistency of interpretation between reviewers.

Reliability and Validity are closely related and often combined to establish confidence in a model. We have not considered reliability in this study, but in the future, this could be established with intra-rater agreement analysis using Pearson's correlation test.

As part of an iterative validation process, we recommend conducting a full validation in line with the procedure defined by Craig et al. (2012) and compare to the validation results achieved in this study. The Hybrid Cognitive Validation Framework was primarily implemented by a researcher aware of the development process and somewhat familiar with the model. We plan to conduct usability assessment and refine the design of the validation framework based on the results. To confirm the utility of the framework, it should be implemented by external research teams looking for a lightweight, initial validation approach for cognitive modelling.

The value of this Hybrid Cognitive Validation Framework comes in the early stages of model development to ensure validation is considered at an early stage and appropriate evidence captured. This framework can be easily and quickly implemented by a small research team and provides a structured validation method to increase confidence in a cognitive model and provide evidence to support validity claims in the early stages of development. This method has been derived from the necessity of the COVID-19 response requiring no direct contact, however we believe that it can have broader application to lower costs and reduce timelines, enabling faster progress in the development of cognitive engineering systems.

Please see the Reference section at the end of the document

End of Paper: Development of a Novel Hybrid Cognitive Model Validation Framework

for Implementation Under COVID-19 Restrictions

Development of a catheterization system AI

The second section of this chapter details the development of an AI to support experts in cardiac catheterization. My purpose here was to understand the complexities of the development of machine agents, and the breadth of options available. There are many categories of AI I considered, from basic statistical models that have been widely used for decades such as linear or logistic regression. More recently, AI has been seen to represent complex learning algorithms that many people view as something of a black-box, providing information without explaining the reasoning behind them to the user. The uncertainty in, and importance of, the definition of AI, is discussed by Monett et al. (2020), consequently, I adopt the established definition of intelligence in the context of AI given by Wang (2008):

> *The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets 'intelligence' as a form of 'relative rationality.' (p.373).*

Therefore, more generally an AI can be classified as any machine capable of intelligence. As AI becomes more complex, there is potentially an increased barrier to trust in the system and can make decision-making less transparent (Mercado et al., 2016). In this stage of the research, I consider machine learning (ML) models as the primary candidates for the AI development as they offer potentially the largest performance gains but also potentially the most significant considerations in terms of human-machine teaming and JCS.

Here, the problem I faced was to identify a specific problem associated with provision of advice on catheterization to practitioners. There are several performance measures in catheter insertion as detailed in the IC3ST (Riga et al, 2011), of these, 'wall-hits' or the number of times the catheter tip hits the side of the blood vessel, is the main candidate for AI-based decision support advice. This is a task currently reliant on expert oversight, which is both expensive and subject to human variation. Objective performance tracking enables surgical progress to be monitored without loss of focus and reinforces learning by enabling real-time awareness and faster correction of mistakes (Barsuk et al., 2009).

To enhance my understanding of AI, and specifically DL-based image classification, I developed a DL model to detect wall-hits on a catheterization task, trained on images obtained from the low-cost catheterization simulator, which I developed for the purposes of this research. I conceptualized a cardiac catheterization decision support system, including the development of requirements for decision support in simulator training and operational angioplasty procedures. This system required selection of a catheter and guidewire combination that best suits a given procedure, patient, and surgeon (Myler, Boucher, Cumberland, & Stertzer, 1990). The deep learning image classifier developed is detailed in the following paper.

In the context of the AI copilot example, this paper relates to developing a understanding of the underlying AI that powers the co-pilot. This particular example is of a deep learning image classifier, but in the example, this could be integration of sensor data or other algorithmic systems required for the specific task. Deep learning was chosen as it is one of the most challenging cases for explainability. Here the XAI framework cues the designer to consider the nature of the AI itself, along with the potential for bias in the data, and AI itself.

Deep Learning for Classification of Wall Hits in Cardiac Catheterization Simulators

Stone P.B., Ganapathy. S., (2022), submitted to the International Journal of Human-computer interaction, under review.

ABSTRACT

Intravascular catheterization is a complex task requiring expert surgeons to train junior doctors. This paper focuses on the development of a deep learning image classification model to assist expert assessment of vascular wall hits in a catheterization simulator. We utilize a transfer learning approach, taking the VGG16 image classification model, and fine-tuning it on wall-hit image data. The retrained VGG-16 classifier achieved a precision of 0.94 and a recall of 0.91, along with an f-1 score of 0.92 on test data from the catheterization simulator. This study demonstrates that vascular wall hits can be detected using a deep learning classification model. These results are in line with the expected performance of the VGG16 model and broader state-of-the-art for image classification. This model shows promise for enhanced catheterization assessment to assist expert analysis or provide objective feedback to trainees.

Background

Intravascular catheterization is a complex, time-critical medical technique that underpins minimally invasive procedures from angiographic imaging to angioplasty and stent fitting. Typically, the procedure involves the placement of a substance or device in the patient’s vascular system by means of a catheter and guidewire combination, introduced via a cannulation site typically via the femoral or subclavian vein, described in detail by Goldman and Pier in 1993. The surgeon manipulates the guidewire and catheter to the desired location monitoring the position within the patient by means of X-ray Radio Fluoroscopy (XRF). Although this is a highly skilled procedure, there is no structured training process and techniques, and skills are typically passed in an ad-hoc manner by supervising surgeons to trainees and junior doctors in their training phase. While this has been an effective means of training vascular surgeons, it makes standardization and performance tracking somewhat difficult and can limit the expansion of best practice within the profession. One way to improve training outcomes would be to implement a standardized training system for early phases of catheterization training.

Trainees in cardiac catheterization program must follow specific steps delineated in the COCATS4 (King et al., 2015) training requirements to gain appropriate experience in the cardiac catheterization lab. COCATS 4 Taskforce 10

(King et al., 2015) outlines a structured, three phase framework for training in cardiac catheterization and is endorsed by the Society for Cardiovascular Angiography and Interventions. The framework outlines milestones in knowledge and skill requirements in a detailed timeline. The ability to perform a Right Coronary Artery (RCA) Catheterization is a requirement of the first phase of COCATS 4 (King et al., 2015), expected to be complete after 24 months of medical training. This procedure is the focus of this study, and we believe supporting early phase training can improve performance. An interactive model involving digital content can support the pedagogical need of teachers to facilitate learning and individualize lessons in a manner superior to traditional, more passive approaches.

Catheterization training is either simulation-based or conducted on cadavers in the initial stages and highly reliant on expert mentors to assess performance. The requirement for expert oversight limits the amount of time available for training, even in the early stages. Barsuk et al. (2009) demonstrated that Cardiac Catheterization simulation training can increase both the skill and self-confidence of trainees. In a subsequent 2010 study Barsuk et al. also investigated the long-term effect of simulation with between 82.4% to 87.1% of trainees maintaining their performance up to one year after training. These findings support our assertion that simulator training can contribute to improved catheterization performance as part of a structured training program such as defined in COCATS 4, Taskforce 10 (King et al., 2015). A key aspect of this training is the assessment of catheterization performance. The “IC3ST”, catheterization assessment framework previously defined by Riga et al. in (2011).

provides an existing baseline for use in catheterization training. While this framework was designed for use in the assessment of robotically assisted catheterization, the performance metrics were also used to compare to human-only performance and are equally applicable to training assessment. An important performance metric defined by Riga et al (2011) is contact with the vessel wall/vessel trauma, also called “wall hits”. Wall hits are defined as events within the procedure where the tip of the catheter or guidewire makes contact with the vessel wall. The procedure requires wall hits to be assessed by the catheterization expert overseeing the procedure and the level of contact is scored on a five-point scale with a successful procedure having “minimal” wall hits and a poor procedure being defined as having “excessive” wall hits. While there are no defined thresholds for “minimal” or “excessive” quantifying the total number is an important part of the assessment task.

The assessment task requires high levels of attention and focus from the expert and wall hits can be brief events, lasting a fraction of a second and can easily be missed. This assessment task itself uses valuable surgical resources, adds to costs and in the case of training introduces performance measures that are largely dependent on the subjective opinion of the assessing expert. We believe there is an opportunity for AI to be utilized both to improve assessment performance in simulated catheterization and increase training time without increasing use of experts. To do this, we propose using the image analysis power of AI, and specifically deep learning, to count wall hits in catheterization simulation, to provide cuing to expert assessors in the traditional training context, or potentially allow trainees to conduct simulated catheterization and receive feedback without the need for expert oversight. While such a system does not replace the rich and nuanced feedback given by experienced practitioners, it does allow trainees more opportunity to practice skills with feedback to enable reinforcement of good performance.

Evaluation of competency is a key element of the COCATS 4 (King et al., 2015) training framework and it is not intended that such a system could or should replace the expert judgement associated with this. The intent of such a system is to support additional learning on top of expert guided instruction where needed and could also be a source of information to support experts in their guidance and decision making. We believe implementing an AI-based hit detection application into a training system could support trainees in the early phases of medical school, without compromising any aspect of the mentor-trainee relationship. This approach has the potential to reduce error and increase objectivity in the assessment of catheterization and along with increased exposure to simulation-based training. This increased catheterization simulation training has the potential to improve performance and associated patient outcomes (Barsuk et al., 2009). While there are many problems that need to be explored in the development of an enhanced catheterization training system with wall hit detection, the ability of the AI to detect a wall hit from an image is fundamental to this concept. Therefore, the key problem we set out to answer in this study is; “Can an AI detect wall hits in a catheterization simulator?”.

Deep learning has shown improved image classification performance over traditional machine learning techniques as discussed by LeCun, Bengio, and Hinton

(2015). Keskar et al. (2016) defined deep learning as a representation-learning method with multiple levels of representation obtained by composing simple but non-linear modules that each transform the representation at one level (starting with the raw input) into a representation at a higher, slightly more abstract level. Deep Learning systems are usually based around a feed-forward deep network or MultiLayer Perceptron (MLP) effectively a deep implementation of Artificial Neural Networks (ANNs), as detailed in “Deep Learning” by Goodfellow, Bengio, and Courville in 2016. The structure of the deep learning network is a feed-forward system that takes an input in the form of a tensor and applies a series of hidden layers made up of one or more neurons each with a weight and activation function defined within it. The output of each neuron then becomes the input for the next neuron in the network. Once an input has passed through the system it is then passed through an output layer. The key to the learning ability of deep networks is the training. The output of the feed forward network is compared to the defined input via a loss function which assesses how well the deep learning model has approximated the input (Goodfellow et al., 2016). The model then implements a process to minimize this loss function value. Typically, Deep Learning systems utilize stochastic gradient descent (SGD) to achieve this implementing the back propagation of error through each layer of the model to determine how much to update the weights of each neuron (Goodfellow et al., 2016). Most deep learning systems also include a learning rate parameter which determines how quickly the SGD algorithm attempts to reach the global or local minima in the loss function. This technique is increasingly showing potential to conduct automated image analysis in a range of applications from facial recognition to assessment of medical imaging.

While the deep learning approach is to minimize the pre-engineering of features, there are decisions to be made when building the system that can maximize the potential for a deep learning system to conduct the selected task. There are many different approaches within deep learning that could potentially be used to automatically detect wall hits from catheter insertion imagery. In this study, we utilize a transfer learning implementation of an existing convolutional neural network (CNN), VGG-16 (Simoyan et al., 2013). This allows a framewise analysis of the video stream and assumes a wall hit can be determined from analysis of individual frames. For the initial study, we focus on the simple case where the wall hit occurs as the movement of the catheter and guidewire is roughly orthogonal to the direction of the field of view of the camera and the intersection with the wall by the catheter or guidewire tip is clearly visible, it is believed both methods should be equally effective. Our research hypothesis is that a fine-tuned version of the VGG-16 image-classification CNN architecture, developed by Simonyan and Zisserman (2013), will be able to classify images with similar precision, recall and f-1 score metrics to the underlying VGG16 network.

A Deep Learning approach for the identification of metastatic breast cancer from image analysis is described by Wang et al. (2016). The authors demonstrate the potential for using transfer learning refinement of the VGG16 model in medical imaging applications. Accuracies of over 97% were achieved on sample classifications of 90 Magnetic Resonance Imaging (MRI) images (Wang et al., 2016). By comparison, the initial application of the VGG16 network was to classify images from the ImageNet dataset and achieved a top 5 accuracy of 92.9% (Keskar et al., 2016), although this was demonstrated on a dataset with considerably more classes. The ImageNet task is more complex than the binary classification of the hit/no hit in the case of catheterization performance, but it could also be argued that the wall-hit classification is a more challenging task as the two classes are highly similar, determined by small differences in small areas of an image. As this is a proof-of-concept study, we expect that the performance of the fine-tuned VGG16 network to be close to the 93.9% accuracy achieved against the ImageNet dataset.

This hypothesis was tested on data collected during the testing of a low-cost catheterization simulator developed by the Human Performance Laboratory at Wright State University. This offers a small but readily useable data set that is free from human-subjects data protection requirements and has expert consideration of “truth” for wall hits time-synced to the video data.

Materials and Methods

To conduct the comparison of the ability of the Deep Learning networks to classify wall hits, suitable baseline data was required to train, validate, and test the networks. In this initial study, a catheterization simulator was used as this enabled fast data generation with the ability to provide ground truth for each wall hit event. The low-cost catheterization simulator developed by the Human Performance Laboratory at Wright State University, shown in Figure 1B.1, was used to generate data in this study.

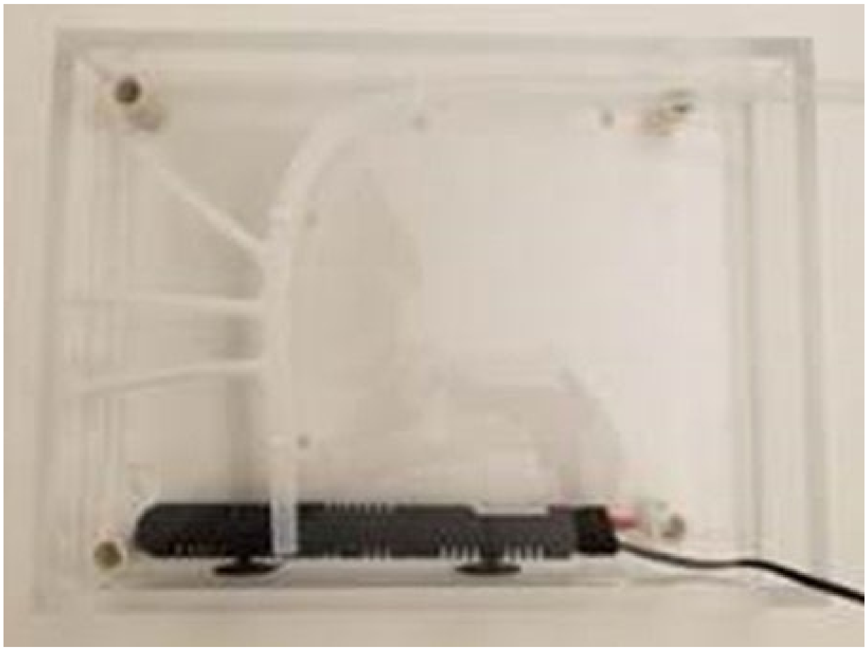


Figure 1B.1. WSU Low-Cost Catheterization Simulator (including heater).

The simulator provides a 16mm catheterization pathway, approximately 450mm in length with three branches to simulate different catheterization locations, with varying branch diameter and bifurcation angle. Clear silicone tubing with a wall thickness of approximately 2mm is used to simulate the vascular system and the whole system is immersed in a 60% glycerin, 40% water mix to simulate the viscosity of human blood (Riga et al., 2011) The mixture is maintained at approximately 97°F to simulate human body temperature. This is particularly important to ensure the catheters are at operating temperature as they are designed to operate inside the human body and temperature variations can affect performance. In addition to the temperature control, catheters used in the simulator during this study were treated with a water-based lubricant to minimize friction between the catheter and the silicone tubing.

Catheter insertion tasks were performed on the Low-cost catheterization simulator, developed at the Wright State University Human Performance Laboratory.

The simulator set-up, including clear silicone tubing allows for direct line-of-sight assessment of wall hits during data collection where the interaction between the catheter and the silicone tubing can clearly be seen from all angles and coupled to the tactile feedback from the catheter without the mechanical damping from the human body. Standard catheterization procedures were followed in the data capture process conducting multiple insertion tasks until a minimum of 20 separate hits were observed for each condition. Hits were not predetermined but were noted when observed imagery was recorded using a fixed webcam directly above the simulator with a Field of Regard (FoR) of approximately 5x3 inches focused on each location in turn. Video imagery was captured via a Hue HD pro webcam using a 1280x720 avi format with mp4 encoding at 30 frames per second. The video was encoded with RGB color and has 3 channels of data per pixel. The images were manually classified prior to the study. The criteria for a wall-hit was any contact between the tip of the catheter and the "vascular structure" of the simulator. This was achieved through framewise analysis of the video capture.

Image capture was conducted at two locations on the simulator. These two locations were associated with typical navigation tasks from a catheterization procedure, specifically curvature and bifurcation navigation tasks. The two locations are shown in Figures 1B.2 and 1B.3.

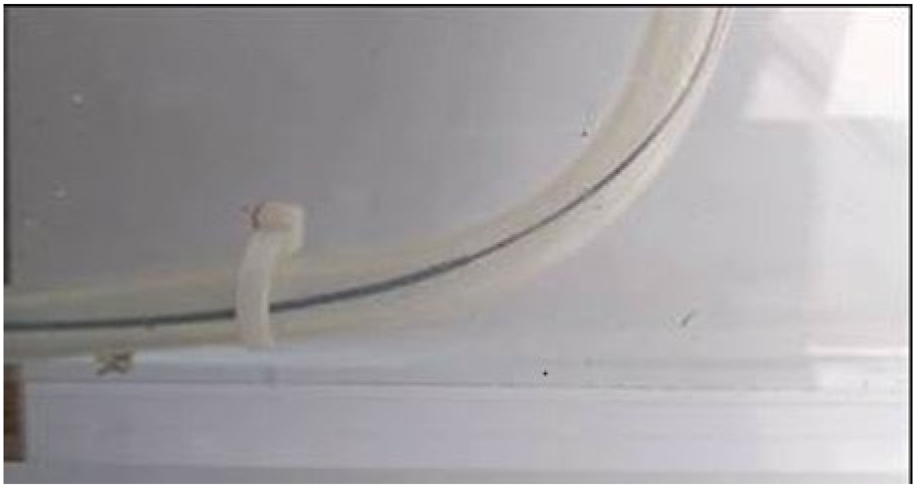


Figure 1B.2. Field of Regard from curvature navigation task.

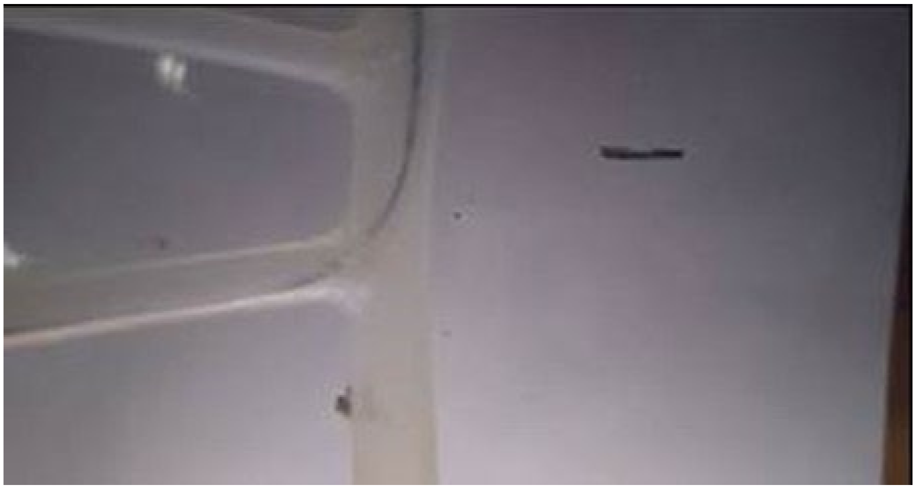


Figure 1B.3. Field of Regard from bifurcation navigation task.

This set-up is not representative of the fluoroscopic imagery used in vascular surgery, lacking the complexity of vascular structures, low contrast and temporal properties associated. Despite these limitations, the experimental set-up is relevant to the proof of concept for a catheterization simulator automatic image assessment tool. a proof-of-concept study to evaluate comparative performance of Deep Learning architectures.

Four cardiac catheters were used including: two variants of the Amplatz-Left catheter (c1, c2), one Multipurpose catheter (c3) and one Judkins-Right catheter (c4). The catheters, shown in Figure 1B.4, were available for testing based on availability of unusable stock and while typical of potential catheters used in training, only represent a small fraction of available catheters. These were used on two different simulator tasks (navigation to two different branches of the simulator). These catheters represent a range of the different tip shapes and sizes available to surgeons but are only a small fraction of the hundreds of designs available.

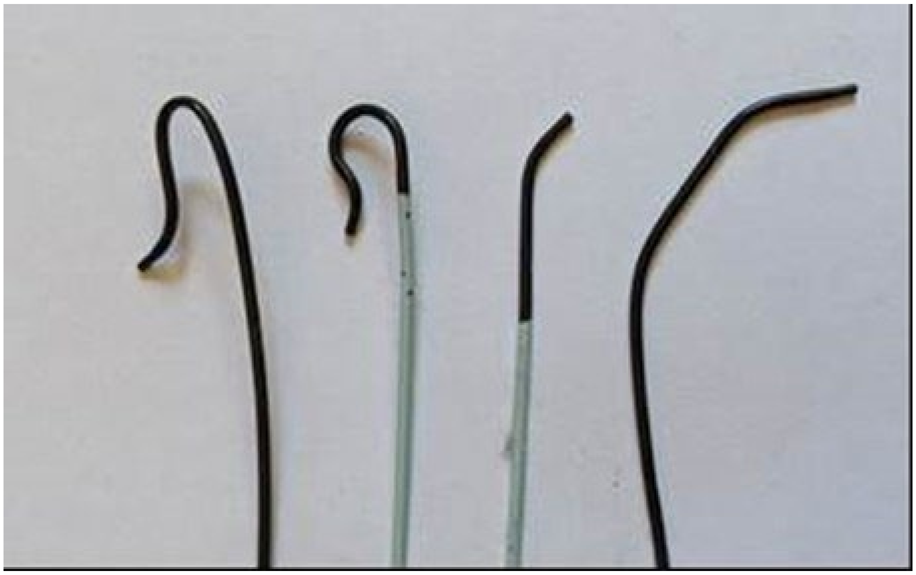


Figure 1B.4. Coronary Catheters used in the study (Judkins Right A(C1), Judkins Right B (C2), Multipurpose (C3), Amplatz Left (C4).

In each case ten videos with no wall hit were collected along with ten videos containing single wall hit events. This resulted in a total of 80 videos split into 40 “no hit” samples and 40 “hit” samples accordingly. This is a relatively low sample size for Deep Learning purposes with typical training sets being made up of hundreds or thousands of samples, if not more. While this is not ideal and could mean resultant

Deep-Learning models do not generalize well, this limitation will be considered in the application of the results of this study.

Image data were collected at two maneuver locations in the simulator for four catheters, resulting in eight experimental conditions. Twenty repetitions were conducted at each condition resulting in 160 images. This was repeated for the hit and no-hit classes to ensure a symmetrical dataset for ease of analysis. In total, 320 images of the catheterization simulator were captured and utilized in this study. The images were randomly split in code to ensure no bias in the split. The data were split into a training set (60%), validation set (20%) and evaluation set (20%). Data were drawn symmetrically from the “hit” and “no hit” classes when generating these data sets resulting in equal numbers of images from each class in all three datasets. The test data were put aside and not used in the design and training phase of model development.

The measures of Precision, or Positive Predictive Value (PPV), Recall or Sensitivity and f1 score are widely used to assess the performance of Machine Learning models and we use them in this study to ensure consistency with standard practice. In addition to these summary metrics, the Receiver Operating Characteristic (ROC) Curve is plotted to show the potential sensitivity of the model to the probability threshold for classification. This metric may also be useful in assessing potential tuning of the model to align with requirements for limiting false negatives or tuning the sensitivity or PPV of the model.

The transfer learning implementation of the VGG16 CNN was based on the approach defined by Rosebrock (2019). The model was built using Python 3 in Jupyter

Notebook on the Anaconda Platform. Full details of the code and Libraries utilized are given in the accompanying code. The most significant libraries used along with the Python distribution were the Keras (Chollet et al., 2013) and Tensorflow (Abadi et al., 2016) libraries used in the development of the Deep Learning Model and the OpenCV library (Bradski & Kaehler, 2008) and FFMpeg (Tomar, 2006). frame extractor used in the preparation of input files for the models.

As described by Rosebrock (2019) the model output layer was initially modified to conduct the first stage of the transfer learning process. In this stage, the ImageNet weights from the VGG16 model are frozen and the output layer was modified so the Fully Connected layer had the same number of outputs as the number of classes in this image classification task, in this case, two. Due to the small size of the dataset, 200 epochs were necessary to ensure training convergence in the development of the model.

In addition to this data augmentation was implemented in the training to compensate for the small dataset. The data augmentation was implemented on the fly for each training epoch in line with the transfer learning method described by Rosebrock (2019) This meant the dataset remained the same size, but training images were adjusted in line with the data augmentation policy in each epoch. The data augmentation characteristics implemented were randomized according to the following rules; rotation range 30°, zoom Range 0.15 or 15%, width shift range 0.2 or 20%, height shift range 0.2 or 20%, shear range 0.15 or 15%, horizontal flip – true

This augmentation policy provides a wide variation in training data for each epoch to ensure the trained model generalizes to unseen data better than if raw data is used. The model can then be trained on this data using the validation dataset to assess the model against while retaining the unseen test data for final evaluation.

The model was then trained on the initial data and resulted in a training accuracy of 77.9% with an associated validation accuracy of 88%. This result showed that even training the output layer in isolation could achieve moderate image classification performance, however this was not comparable to the VGG16 performance on the ImageNet dataset. To achieve this, the VGG16 model weights were unlocked and the whole model was retrained on the Catheterization dataset, again over 200 epochs.

This resulted in training performance similar to the image classification performance of VGG16 on the ImageNet dataset. An accuracy of 96% was achieved on both the training and validation datasets, which compares well to state-of-the-art and transfer learning implementations of VGG16 in healthcare applications. As training and validation accuracies were in line with expectations, the model was believed to have acceptable image classification performance and the model evaluation phase was conducted against the unseen test data set aside at the beginning of the model development process.

Results

The transfer learning fine-tuned variant of the VGG16 CNN was assessed for image classification performance against the catheterization wall-hit data set. This image classification task was to identify which of two classes (hit or no hit) an image belonged to. The test dataset consisted of 64 images, equally split between “hit” and “no hit” classes. The raw image classification results are presented in the confusion matrix shown in Table 1B.1.

Table 1B.1. Confusion Matrix of catheterization Image classification.

N=64	Predict Hit	Predict No Hit	Total
Actual Hit	30	2	32
Actual No Hit	3	29	32
Total	33	31	64

This shows the successful training performance generalized well to the test data with a 93.8% True Positive rate and a 90.6% true negative rate. This translates to a Precision, or Positive Predictive Value (PPV) of 90.9% and a Recall, or sensitivity value of 93.8%. The f1 score for the image classification was 92.0% overall. The raw probability output from the model was used to generate the receiver operating characteristic (ROC) Curve for the catheterization image classifier. The ROC Curve for the image classifier is shown in figure 1B.5.

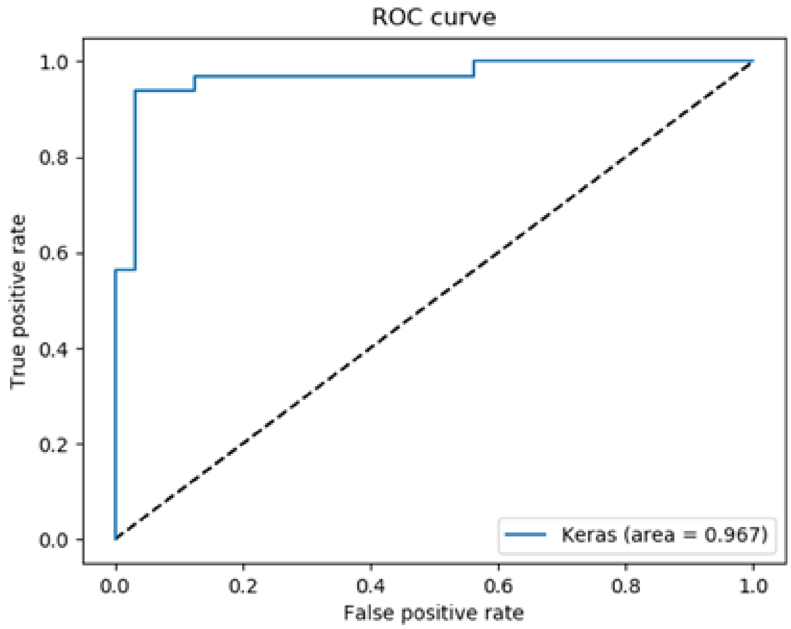


Figure 1B.5. ROC Curve for the wall-hit image classifier.

This shows catheterization classification model can achieve high true positive rates with a low false positive rate with a range of thresholds for classification probability.

Discussion and Conclusion

This initial proof of concept study, while showing good potential for wall hit classification in a catheterization simulator had several limitations to note. The catheters used in this study were not specifically selected based on their properties or representativeness but available on an opportunity basis. Although representative of a range of catheters used in coronary catheterization are potentially not those that would be used on the simulated task. The low-cost simulator is designed to replicate an RCA catheterization procedure and as such, future trials should utilize catheters which are more appropriate to this procedure.

Data used in this study were collected by the research team for sole use in this study. This was necessitated by time constraints and limitations of access to medical professionals associated with the COVID19 outbreak during 2020. The study should be replicated using trainee surgeons utilizing the low-cost catheterization simulator where their focus is the successful completion of the task, and they are unaware of the need to collect wall hit data. This will reduce the potential for bias in the study and ensure data captured is representative of the proposed application.

The "hits" were recorded with the catheter tip moving in the x-y plane of the field of regard. Future work should cover wall hits in any direction, which may be less easy to classify. The transfer learning implementation of the VGG16 network is a CNN based on a framewise data from the video. This classification technique was shown to be effective in the proof-of-concept data, but it is believed a 3D CNN or Recurrent Neural Network (RNN) solution may be better able to classify more subtle wall hits or those that occur along the z-axis of the field of regard or viewing axis. These solutions are sensitive to time data, which in the case of video means being able to analyze data between frames as well as within them. This could be critical to classifying wall hits that are not apparent from a single frame.

Deep learning and neural network implementations on visual data have the ability to do more than predict or classify data. Increasingly the output can be visualized and implementations of image segmentation in Neural Networks can be used to visually identify areas of interest in an image. This technique could be used to not only classify the wall hit but alert the user as to the location of the hit. This provides visual confirmation that increases trust in the system but can also highlight historical data or trends in wall hit location to provide a more engaging feedback mechanism in the training context.

This study demonstrates that Deep Learning can be used to detect wall hits in catheterization simulation, this information can be used to create training methods in the future to improve expert assessment or trainee performance. The image classification system utilized transfer learning on the baseline VGG-16 model (Simoyan et al., 2013) and achieved results in line with similar implementations both in medical image classification and the original ImageNet task. We conclude that the catheterization classification model was able to classify reference images taken from a cardiovascular training simulation as either containing a wall hit or not containing a wall hit. The high Precision, Recall and F1 Score associated with the catheter wall hit classifier indicate that this method could successfully be used to augment existing training systems and provided objective evidence of wall hits. This could underpin the objective metrics of an enhanced training system could be integrated into COCATS 4, Taskforce 10 (King et al., 2015), phase 1 training to provide objective performance feedback and aid expert evaluation, helping to shape training and development goals.

Please see the Reference section at the end of the document

End of Paper: Deep Learning for Classification of Wall Hits in Cardiac Catheterization Simulators

Chapter 2 – An Initial Design Thinking Framework for XAI in TCS

In chapter one, I consider cognitive modelling and the application of AI in the development of expert systems to a specific problem, whereas here I focus on a more generalizable approach to the design and development of AI in human-machine systems. In this chapter, I apply the principles/findings from chapter 1 to the development of user-experience and user-interface elements of a system, specifically in the context of explainability in AI systems, establishing XAI as the key design problem and the subject of the generalized approach.

Explainable Artificial Intelligence

This paper focuses specifically on Explainability in AI systems to improve human-machine teaming. To this end, one of the first questions that needs to be asked is 'what is XAI? In the introduction to this document, the need for XAI is discussed in some depth but it is also important to understand exactly what is meant by XAI. As XAI is still a developing area of research there are a range of definitions that require consideration. Hagras, (2018) defines fives aspects of explainability: Transparency, Causality, Bias, Fairness and Safety. Key among these is transparency, which is the means to provide information on Causality, Bias, Fairness and Safety.

Indeed, the concept of transparency is sometimes used interchangeably with explainability and interpretability. Doshi-Velez Kim (2017) suggest that to interpret is to explain or to present information in understandable terms, where Montavon, Samek and Muller (2018) define interpretation as the mapping of an abstract

concept into a domain more readily made-sense-of by humans. The difference in these definitions is subtle but important – this can be explained in the terms of human languages – where explaining a statement in each language involves rewording a statement where the individual words are understood but the meaning of the sentence is not, versus interpreting a statement in a foreign language to a language known by the reader. This interpreted statement may still require further explanation.

Given these definitions, interpretability can be considered a step along the way to explainability. For deep learning to be explainable, there may need to be an interpreting of the model rationale from terms that the deep learning designer might understand to those more easily understood by the user. This stage of the process is important no matter which definition of interpretability and explainability are used. Explainability therefore combines the need for both transparency and interpretability. Specific instances of XAI will need transparency measures and if appropriate, the transparency information may require interpretation or translation to terms that the end user may understand.

Structured design methods

There are many approaches to structure design problems that seek to give engineers and designers the tools to address these complex problems. Failure Mode Effect Analysis (FMEA) is a widely used and adapted method for structured problem analysis. Design FMEA (i.e., DFMEA) adapts this, establishing eliminating failure as a purpose of design (Bueno, et al. 2020). The theory of the resolution of invention-related tasks, known by its Russian Acronym 'TRIZ', is a design framework that focuses on the assumption that most complex design problems require a trade-off between competing requirements (Al'tshuller, 1999). Design Thinking is an approach that covers the design problem from understanding of the problem to testing and highlights the importance of the mindset of the designer and fast iteration of solutions. The Kembel (2009) adaptation of design thinking is noteworthy as it specifically requires designers to empathize with users, and those the solution may impact. These approaches are not exhaustive but give an idea of the breadth of approaches that can be adopted in structuring the design process.

While these methods provide structure to the design process, they are typically iterative and non-linear (Kumar, 2012). Selection of a suitable design framework can also be a complex problem. It is important to understand the strengths and weakness of the framework and the ways in which the strengths complement the design problem at hand. The problem of integration of AI transparency in time-critical decision support is unpredictable and requires the ability to understand specifics of human-machine teaming and the impacts of modifying existing relationships in a given task. In addition to this, it is necessary to be able to iterate quickly and maintain a flexible approach. For these reasons we selected the Design Thinking process defined by Kembel (2009) as the most suitable candidate as it specifically addresses development of a deep understanding of the problem through empathy with stakeholders.

Situational Awareness

The way in which humans frame problems impacts the decisions they make to solve those problems. Good decision-making is highly dependent on SA, and time-critical decision making is particularly dependent on maintaining SA and, in turn, can be associated with diminished SA when tasks become unpredictable, or user resources become saturated (Endsley, 1995). Expert decision making in TCS, particularly complex systems, is described by Naturalistic Decision models like Klein's (1993) Recognition-Primed Decision (RPD) model. The

The importance of SA in TCS and the relationship to technological systems is explained by the model of situated cognition (Shattuck & Miller, 2006). This model represents how technological systems transfer information to humans and the perceptual and cognitive processes that result in a state of situated cognition or SA, see Figure 2.1.

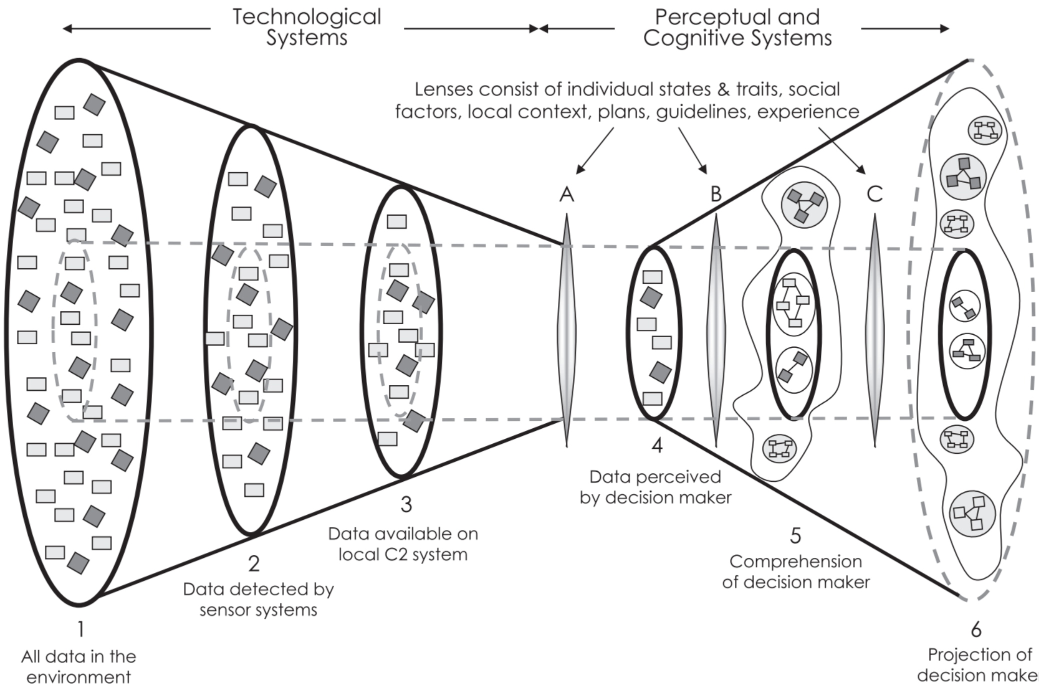

Technological Systems
Perceptual and Cognitive Systems
Lenses consist of individual states & traits, social factors, local context, plans, guidelines, experience
A
B
C
1
All data in the environment
2
Data detected by sensor systems
3
Data available on local C2 system
4
Data perceived by decision maker
5
Comprehension of decision maker
6
Projection of decision maker

Figure 2.1 A model of situated cognition (Shattuck & Miller, 2004)

This model shows how HMTs work to achieve the three levels of situational awareness in users; perception, comprehension, and projection. But also, how there are decisions made by technological systems that limit the focus of the user – through the lenses on the left-hand side of the model. This model is augmented to become a dynamic model of situated cognition, including feedback loops, that represent the way experts modify their view of the world based on experience and the evidence presented to them, see Figure 2.2.

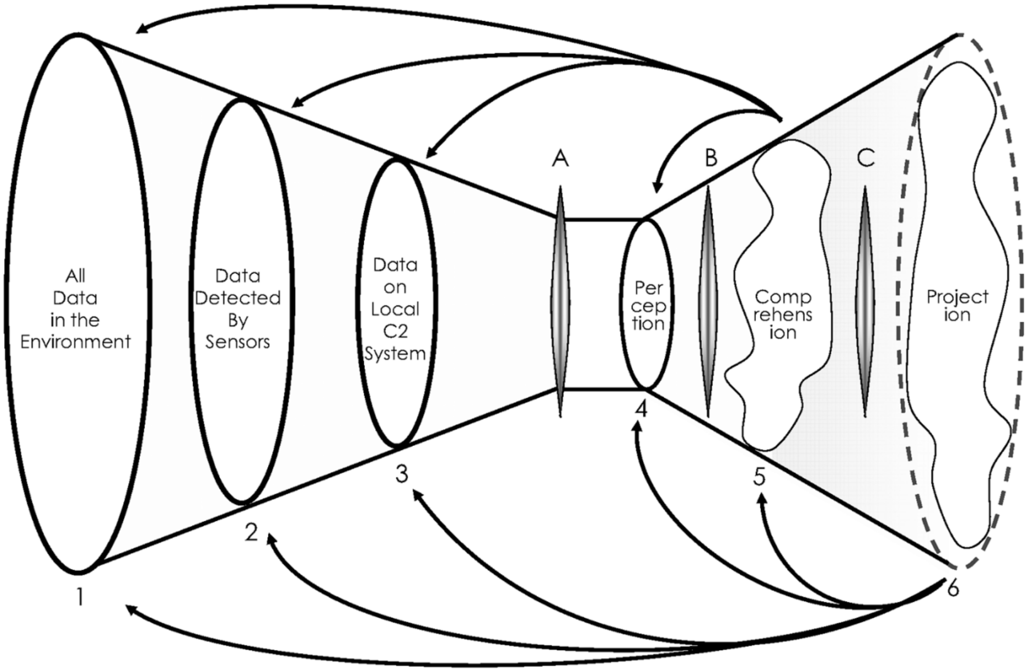


Figure 3. Feedback Loops in the Dynamic Model of Situated Cognition
Copyright © 2003 N. L. Miller and L. G. Shattuck. Reprinted with permission

Figure 2.2. A Dynamic Model of Situated cognition (Shattuck and Miller, 2006)

In addition to the representation of SA in TCS, this model can be adapted to represent XAI in TCS. The right-hand-side of the model remains the same, and confirms the importance of SA, where the left-hand-side of the model becomes a representation of the AI. The feedback loops represent trust and the sizes of the perceptual lenses, particularly lens A, are modified by workload – the lower the workload of cognitive demands, the more the user can absorb additional information.

This explains how XAI, through transparency along with, trust and Workload contribute to SA and correspondingly can lead to reduced or imperfect SA if they are compromised. Trust acts as a feedback loop and is modulated by the accuracy of the AI system and the ability to calibrate trust provided by the XAI. Workload modulates the amount of information available to the user but the amount of XAI information displayed can also limit information transfer. The tradeoff can be optimized by understanding how much information can be transferred for a given level of workload and is dependent on the user's cognitive ability, their trust, and the requirement for SA in any given task. If SA becomes compromised, this can lead to imperfect SA. The problem of imperfect SA relates to how decision makers can recognize problems in SA and seek to correct them (Middleton, 2010). While there can be additional cognitive load in interpreting transparency information, either reducing the potential for imperfect SA or providing intuitive means to identify or predict uncertainty in the decision context can allow decision makers to quickly identify mitigation strategies (Middleton, 2014). Intelligent systems, capable of predicting task requirements and tracking the expected knowledge states of decision makers can enable DSS to predict when imperfect SA might arise and potentially optimize information provision to reduce workload impact. (Betts, 2005).

Time-Critical Tasks

Time critical systems were chosen as the focus of this research as they represent a complex implementation of XAI and transparency. Not only is there the positive impact of XAI information on trust and task performance, but there is also a need to consider the negative impact of additional information, and the cognitive processing required from a user. Where the user is subject to high temporal demand and high workload, with potentially minimal cognitive resource available to attend to new information, this additional burden may mean XAI having unintended negative consequences, which designers and developers need to be aware of.

In the following paper titled ‘Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support’ (Stone et al., 2022), published in the International journal of Human-Computer Interaction, I detail the development of the design thinking framework, and further explore the importance of trust, workload, and SA in developing XAI and transparent AI systems.

In the context of the AI copilot example, this stage establishes the five-stage design thinking approach as the basis for the XAI design framework, integrating the deep understanding of the user and the AI established in chapter one into the first of these five stages – empathy. Furthermore, this chapter establishes specific tools to enable the definition of design goals for XAI. In the context of the copilot example, this would give the prospective designer of the system, a guide as to the scope of explainability and the potential areas that solutions may be developed, without requiring XAI expertise in addition to their ability to design flight systems for drones. This element of the framework aims to facilitate development of design goals and early concept development without introducing specific goals that might be too restrictive given the broad application of XAI and the task-specific nature of potential solutions.

Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support

Stone, P. B, Jessup, S. A., Ganapathy, S., Harel, A. (2022). Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support International Journal of Human-Computer Interaction (IJHCI) Special Issue on Transparent Human-Agent Communications.

ABSTRACT

The integration of artificial intelligence transparency in time-critical decision support is complex and requires consideration of the impact on human-machine teaming. The relationships between transparency, trust, workload, and situational awareness are key to understanding this impact on performance. We detail the development of a novel design framework for transparency integration in Decision Support Systems. We selected the design thinking approach as the baseline for our framework as this focuses on developing empathy with users and rapid design iteration. We adapted this framework by introducing the concept of empathy for both human and machine agents. In this situation, ‘empathy’ provides a deep understanding of the model, its purpose, and the underlying data for AI. We developed a structured problem definition focused on understanding the relationships between constructs and established solution themes to guide the designer. We demonstrate this transparency integration framework on a Transfer of Care Decision Support System.

Introduction

As Artificial Intelligence (AI) systems become more widespread and integrated in our lives, the need for effective and efficient Human-Machine Teaming (HMT) is increasingly important. This paper describes the development of a design framework for the integration of AI transparency measures into time-critical decision support. We introduce this subject by discussing the importance of transparency in AI, and its impact on HMT. Next, we analyze the specific contributing factors to transparency in AI such as trust, workload, and situational awareness, and the tools for improving HMT in decision support in time-critical situations. Finally, we introduce the concept of design frameworks as a means to solve the integration problem. In the subsequent sections, we detail the development of a specific design framework and a preliminary implementation conducted on an existing decision support interface.

As AI-based systems become increasingly complex and abstract from traditional decision making, transparency to users becomes more important and harder to establish, posing conceptual, legal, and technological challenges (Wachter et al., 2017). So called “black-box” systems can lead to reduced trust, a lack of understanding of the rationale behind the problem, or potentially an over-reliance on high-level advice (Kim et al., 2020). In this paper, we adopt Chen et al. (2014)’s definition of transparency: “…the quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent’s intent, performance, future plans, and reasoning process” (p. 2). As for the definition of machine/computer/AI agent, throughout this paper we combine two complementary definitions: firstly, as possessing “autonomy, observation of the environment, action upon the environment, and activity toward achieving certain goals” (Mercado et al., 2016, p. 401); and secondly, “as an entity that runs by computerized algorithms and that interacts with humans” (Jessup et al., 2019, p. 482). These two definitions cover the meaning of an agent in terms of external and internal characteristics, respectively. Mercado et al. (2016) highlight the importance of transparency, demonstrating a significant multivariate improvement in operator performance through the implementation of transparency in AI.

AI-based Time-Critical Decision Support Systems

In the current paper, we focus on Decision Support Systems (DSS) as a prime example of a domain that can benefit immensely from increased transparency. Decision support systems allow decision makers to utilize data and models to reduce errors and workload and improve Situational Awareness (SA). Integrating AI into DSS, therefore, has the potential to make them more responsive, personal, and ultimately useful, with the potential to overcome equivocality in decision making, lowering user workload, and improving performance (Jarrahi, 2018; Woods, 1985). However, DSS often use Machine Learning (ML) algorithms (e.g., agent-based DSS, genetic algorithms, deep learning algorithms), which tend to become less transparent and less interpretable as they become more complex, leading to lower trust (Strobel, 2019). The major flaw with deep learning is that, despite its power to analyze large data and form complex representations and predictions, there is often little clarity provided to users about how the output is determined (Topol, 2019). Moreover, ML algorithms rely on big data pulled from areas such as social media, organizations, and census records, just to name a few. Because biases are present in humans, and consequently the data they generate, these biases can influence the information generated. For example, Datta et al. (2015) found that based on users' genders, ads for higher paying jobs were shown to more males than females through Google. Perhaps one of the more of the more memorable examples of bias in AI was from Microsoft's Twitter chatbot, Tay, which was shut down for harassing other users and posting tweets endorsing Nazi ideology (Johnson, 2020). Such biases can perpetuate inequality and discrimination in our society, leading to distrust in AI (Thelisson et al., 2017). In other words, AI and automation have wider societal implications that can affect trust and the biases in human agents, and thus, increased transparency is key to improving communication, increasing trust, and facilitating effective HMT (Hoffman et al., 2002; Mercado et al., 2016).

The Effects of Transparency on Trust, Workload, and Situational Awareness

Time-critical tasks increase the need for human operators to rapidly process information and make decisions, potentially resulting in errors and delays (Horowitz & Barry, 2013). Thus, although it is important to consider the strengths and weakness of machine agents, it is also important to understand the human user. Consideration of both human users and machine agents can improve collaboration and performance by reducing errors due to poor communication and task assignment (Woods, 1995), as well as improving HMT (Hollnagel & Woods, 1983). Shared mental models can enable anticipation of other team members actions but depend on the ability of the machine agent to collect enough information to predict the user state (Fan & Yen, 2010). Specifically, advanced AI systems need to identify human cognitive processes through pattern recognition capabilities and real-time feedback to predict changes in their cognitive model before they are implemented to provide optimal timing of transparency advice (LeCun et al., 2015). Transparency in AI also impacts workload (Helldin, 2014) and the legitimacy of information (de Fine Licht, 2011), which can in turn affect trust (see below) which can all impact primary task performance, particularly in time-critical situations. Delays and failures in making time-critical decisions are often expensive, can affect system performance, and may even cost human lives (Sheridan, 1997). Environments demanding time-critical DSS inherently require minimal user interaction with the system to maximize attention on the primary task (Horvitz & Barry, 2013). Therefore, there is an important trade-off in the integration of transparency in AI; although additional transparency information is needed to increase trust, this may also increase workload and reduce SA in situations when users already have minimal mental capacity to spare. Thus, a better understanding of both trust and workload is required in order to better understand the integration transparency in AI.

Trust

Transparency is key to establishing trust in both human and human-machine teams, to prevent negative outcomes and create accountable systems (de Fine Licht & de Fine Licht, 2020). Trust is a willingness to be vulnerable to another, without the capability to monitor their actions. When there is an element of risk and the trustor (the one who is trusting) needs to interact with another party or entity (the trustee) to accomplish a task or foster a relationship for long-term interactions, trust is essential (Mayer et al., 1995). Trust is a social construct that has been identified as important for not just the development and lifetime of interpersonal relationships (Jones & George, 1998; Lewicki et al., 2006; Simpson, 2007), but also for interactions with non-human entities such as machines (Muir, 1987), robots (Hancock et al., 2011), automation (Schaefer et al., 2016), and AI (Glikson & Woolley, 2020).

Malle and Ullman (2021) reviewed numerous definitions of trust from interpersonal, business, and automation domains. They found that overall, trust is comprised of two factors: performance trust and moral trust. Performance trust contains facets of competence and reliability, whereas moral trust contains facets of sincerity, benevolence, and integrity. Compared to human-human or human-robot interactions, human-automation interactions usually involve non-social tasks (Malle & Ullman, 2021), which do not require moral trust. As such, performance is the main factor related to the automation itself that influences trust (Hoff & Bashir, 2015; Hoffman et al., 2013; Malle & Ullman, 2021). Consequently, most trust definitions in the automation literature do not reference the automation's moral characteristics. Indeed, Lee and See (2004) define trust in automation as an attitude an agent will help the trustor achieve their goal in times of vulnerability and uncertainty.

Trust is a process with several components that influence some sort of outcome. Antecedents to trust (characteristics of the trustor and characteristics of the trustee), trust intentions (willingness to be vulnerable), risk-taking behaviors (behavioral trust, which has also been referred to as reliance; Lee & See, 2004), and perceived risk, which moderates the relationship between trust intentions and behavioral trust, are all components of the trust process (see Figure 2A.1). Behavioral trust has also been referred to as reliance in the trust in automation literature (Lee & See, 2004). Trust and reliance in machines are subtly different concepts, and sometimes confused with each other. Trust is an intention to be vulnerable, whereas reliance is the behavioral outcome related to trust (i.e., being willing to drive in an automated car versus getting behind the wheel and going for a drive). In developing a wider understanding of trust, it is important to define the characteristics of automation and the user. It is also important to point out that human-human teaming often entails two-way, reciprocal trust. However, in human-machine teaming, typically the human will be the trustor and the AI or machine agent will be the trustee. Hence it is important to understand not only trust, but also trust calibration.

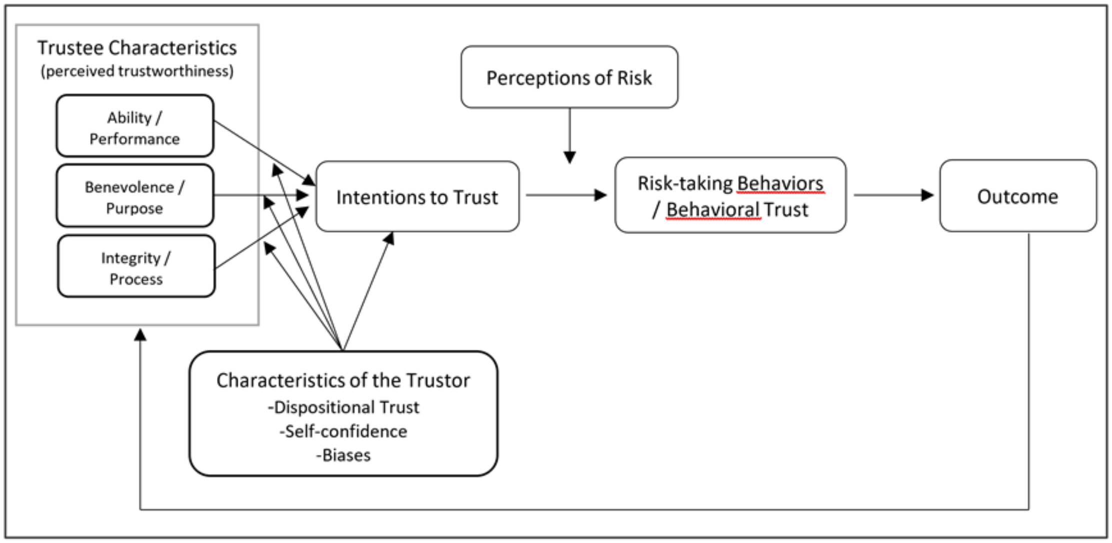


Figure 2A.1. Trust Process Model

Trust Process model based on Mayer et al.'s (1995) theoretical conceptualization of the trust process, and Lee and See's (2004) conceptualization of the three trustworthiness factors related to trust in automation.

Perceived trustworthiness of automation is comprised of three factors: a) performance, b) purpose, and c) process (Lee & See, 2004), which were derived from Mayer and colleague's (1995) factors of perceived trustworthiness in humans (i.e., ability, benevolence, and integrity). In support of Malle and Ullman's (2021) delineation of performance and moral trust, one can see that perceived trustworthiness of automation factors are strongly related to the performance of the referent, whereas Mayer et al.'s factors of interpersonal trustworthiness contain both performance and moral trust. Performance conveys information to the user about what the system does, as well as the boundaries of its capabilities and limitations. The reliability of the agent, as well as the predictability and severity of errors will influence users' perceptions about the agent's performance (Lewis et al., 2018). Purpose refers to why the system was created (i.e., for what tasks will it be responsible), and whether it actually performs those functions. Process describes how the agent will operate. Information about the agent's reasoning, and transparency of the algorithms are important for users to glean understanding about how the agent functions. All three of these factors can increase transparency of the system for users, depending on how much or how little information is relayed because it provides users with knowledge about what, why, and how well functions are being performed. Indeed, researchers have empirically demonstrated that increased transparency leads to increased trust in a system (Lyons et al., 2017). Hengstler et al. (2016) conducted a case study on the factors that influenced trust in applied AI (e.g., autonomous vehicles and medical assistance devices), and identified factors that correspond to Lee and See's (2004) proposed trustworthiness factors. Hengstler and colleagues found that factors such as data security and privacy were related to the performance of the AI, cognitive compatibility (i.e., understandable algorithms) and usability aided in understanding the process of the AI, and defined context (i.e., the environment or task in which the AI was to be implemented) and design (i.e., the feedback provided to users) were used to ascertain information about the AI's purpose. The more information users have regarding the capabilities and limitations of an agent, the better they can calibrate their trust in that agent.

Trust calibration

This refers to "the correspondence between a person's trust in the automation and the automation's capabilities" (Lee & See, 2004, p. 55). One can over-trust, which can lead to compliance and a lack of SA regarding the system (also referred to as misuse; Parasuraman & Riley, 1997). Conversely, one can under-trust (i.e., disuse; Parasuraman

& Riley, 1997), which can lead to inefficiency in performance and increased workload for the user. Properly calibrated trust allows the user to appropriately monitor the system (depending on the level of automation; Parasuraman et al., 2000; Sheridan & Verplank, 1978), while also benefiting from the automation performing tasks in tandem with the user to decrease the user's workload and improve performance (Balfe et al., 2015; Lewis et al., 2018).

Just as characteristics of automation can influence trust, so can characteristics of the user. These characteristics include but are not limited to dispositional trust, self-confidence, and biases. Dispositional trust is the general tendency and willingness for humans to trust others (Mayer et al., 1995), and influences how users will trust others or agents in novel situations, prior to any knowledge about the referent (Lee & See, 2004; Mayer et al., 1995). Another individual difference that can influence trust and reliance is self-confidence (Lee & Moray, 1994; Lee & See, 2004). Empirical evidence has shown that when users have a high level of trust in a system and low confidence in their own abilities, users opt for automatic over manual control to aid in task completion (de Vries et al., 2003). Similarly, biases users have about automation can influence user interactions with automation. Though biases are not inherently bad or good, they do influence people's judgment and decision making (Haselton et al., 2015; Tversky & Kahneman, 1974). Perfect automation schema (Dzindolet et al., 2002; Merritt et al., 2015) is the belief that automation should perform without errors (i.e., high expectations), and if/when the automation fails, it is no longer useful (i.e., all-or-none thinking), which leads to disuse, inefficiency in performance, and increased workload for the user (Dzindolet et al., 2003; Parasuraman & Riley, 1997). Researchers have found that high expectations are positively correlated with trust, whereas all-or-none thinking is negatively related to trust (Lyons et al., 2020). Designers can help to combat this bias by increasing transparency of the automation. Indeed, Dzindolet and colleagues (2003) found that providing participants with an explanation of why the automation failed, increased both trust in, and reliance on, the automation. Informing users of limits, capabilities, reliability, and points where failure might occur will help the user to set realistic expectations and be prepared to intervene, if necessary.

Another bias that influences reliance is automation bias. Automation bias is the tendency of users to over-rely on recommendations from automation, forgoing cognitive processing, even the information received may be contradictory (Mosier & Skitka, 1996). Automation bias most likely occurs when users are trying to conserve mental resources during instances of high workload, during complex, time-critical situations (e.g., command and control), when the operator receives confusing information from the agent, or the user is not properly trained (Cummings, 2004; Goddard et al., 2012). Consequently, automation bias can decrease SA and lead to complacency, possibly leading to catastrophic issues in areas such as patient safety (Schulz et al., 2016) and aviation (Jones & Endsley, 1996). In order to reduce instances of automation bias, researchers have recommended training, providing users with the agent's reasoning process, and determining an appropriate level of automation for the agent to ensure the human remains in-the-loop enough so that SA is maintained (Cummings, 2004; Goddard et al., 2012). Additional factors related to the environment or situation, such as SA, workload, and time constraints can also influence trust and reliance (Lee & See, 2004; Lewis et al., 2018).

Mental Workload and Situational Awareness

Although trust is perhaps the construct most closely associated with transparency, workload is also an important consideration. The impact of transparency on workload is complex and dependent on the nature of the task, the experience of the individuals involved, and any parallel tasks; transparency has the potential to increase, as well as decrease, workload (Helldin, 2014). This increase can be minimized by effective, efficient display of information (Chen & Barnes, 2013) and increased trust and collaborative HMT by contributing to a more intuitive, information dynamic between human and machine agents (Mercado et al., 2016).

The negative impacts of poor implementation of automation can be manifested in many ways, all with significant effects on workload and task performance (Parasuraman & Riley, 1997). When automation acts outside intended parameters and fails to respond to human commands, it can lead to catastrophic failure. Introducing transparency aims to reduce potential for this extreme failure but the potential should not be ignored, particularly in safety critical applications. Recovery from these types of failure can present steep increases in workload or severely impact task performance (Sarter et al., 1997). A more frequent but less well understood problem is when automation behaves as intended but human collaboration is ineffective, causes additional workload (Sarter et al., 1997), or leads to misuse and abuse, which can result in negative consequences because of human error (Parasuraman & Riley, 1997) or automation bias. As with catastrophic failure, these complications can lead to additional workload, or resources required to recover, and diminished SA.

The way in which humans frame problems, impacts the decisions they make to solve those problems. Good decision-making is highly dependent on SA, and time-critical decision making is particularly dependent on maintaining SA and, in turn, can be associated with diminished SA when tasks become unpredictable, or user resources become saturated (Endsley, 1995).

The Structured Design Approach

The problem of AI transparency integration in time-critical decision making is complex with many inter-related constructs. The impact on HMT may be difficult to predict and as such presents designers with a problem. Concepts may behave as expected in certain circumstances but small variations in the environment or task can have large impacts on performance. Solving complex innovation problems requires an in-depth knowledge of the task, the user, and the systems with which they interact. Great innovation can be hard to achieve and requires organizations to develop systems and cultures to support innovation (Kumar, 2012). Although structured design methods provide order to the design process, selection of a suitable design framework can be a complex problem (Kumar, 2012). We suggest that adopting a structured design process that explicitly guides the designer to consider these problems and developing it into a framework that provides tools and design goals can help ensure that transparency issues are considered in the integration of AI in time-critical DSS.

Framework Rationale

The key question designer should ask in this development process is: how can HMT be improved by the integration of transparency in time-critical decision support? To assist the designer in this task, we developed a novel design framework which directly tackles some of the aforementioned problems with HMTs containing an AI-based agent and increase transparency. The rationale of this framework is to apply the design thinking process to enable effective HMT that, in turn, allows better human performance, user experience, and task outcomes. The fast, iterative nature of design thinking is ideal for framing problems to develop understanding of paradoxes and inherent trade spaces in design (Dorst, 2011). We map specific concepts relating to transparency in time-critical decision support into this process to provide a framework for integration of transparency. Specifically, this framework can then be used as a tool for the designer to structure the key elements of the transparency integration process. We believe this framework reduces the potential for design error and speeds up the integration of transparency measures into time-critical decision support applications.

Objective

The objective of our study is to develop a design framework for the integration of AI transparency into time-critical decision support systems. To achieve this, we aim to transform the transparency considerations into additional, context-specific concepts and design strategies to underpin the framework. We will provide structure by combining these design concepts and establishing specific tasks and goals. In addition to the generation of this framework, we conducted an initial implementation of the integration of transparency measures in a previously developed prototype DSS for Transfer of Care (ToC) (Stone 2019).

Framework Development

Rather than develop an entirely new design framework, one of the key innovations in our approach is to map transparency considerations on to an existing, proven design framework. After consideration of several alternatives, the design thinking approach defined by Kembel (2009) was selected as the most suitable candidate as it is a widely used approach to design that covers understanding of the problem through to testing. Design thinking highlights the importance of the mindset of the designer, open and honest exchange of ideas and fast iteration of solutions and specifically addresses development of a deep understanding of the problem through empathy with stakeholders. It is often a quick process aimed at rapidly developing and testing innovative solutions to difficult problems (Kembel, 2009). In addition, the design thinking approach is highly flexible and adaptable to a wide range of design problems but offers a proven concept that designers understand (Thoring & Müller, 2011). The design thinking approach defined by Kembel (2009) differs from similar methods as it establishes empathy as the first of the five stages. The five stages of design thinking process as defined by Kembel are shown in Figure 2A.2.

Empathy
Define
Ideate
Prototype
Test

Figure 2A.2. Five-Stage Design Thinking Process

Note. Figure above was created by authors to illustrate the five-stage design thinking process outlined by Kembel (2009).

We adapted this framework for the design and implementation of transparency measures into time-critical decision support systems by considering each phase and defining the associated sub-tasks and information specific to the challenge of integrating transparency in a DSS. The impact and relationships between trust, transparency, workload, and SA are considered throughout the development to maintain the importance of these concepts in the mind of the designer. The output of this development process are the goals and tasks, which are summarized in Table 2A.1. Next, we detail the stages of the development approach in turn.

Table 2A.1. Design Framework Task Summary

Design Stage	Framework goals	Framework tasks
1 Empathy	• Develop a deep understanding of human and machine stakeholders, their challenges, limitations, and requirements in the context of decision support transparency. • Understand the potential for bias in both data and human users.	**Task 1A** - Task analysis: Conduct a hierarchical task analysis (HTA), including open questions bearing in mind the requirements for presence, open-mindedness, and lack of judgment. **Task 1B –** User analysis**:** Conduct a user analysis to understand the characteristics, motivations, and background of the user population. **Task 1C** - System analysis: Research the data provision, data requirements, and the impact of missing or incorrect data. Additionally, researching the underlying model and how it handles both existing data and data collection are important during this task.
2 Define	• Establish problem definition with respect to the level of transparency, transparency types, and testable design requirements, • Formalize the relationship between transparency, trust, workload, and SA in the context of the specific time-critical decision support task.	**Task 2A** –Transparency task definition: Define specific questions formulated from the understanding of the nature of transparency integration. These questions ensure that the framework guides the designer to think of the key drivers and potential impacts in the Formative stages of the project. Specific questions should include: • What are the specific issues relating to transparency between human and machine agents in this specific context? • What are the existing or future impacts of a lack of transparency? • How can these transparency implementations improve outcomes? • How might poorly implemented transparency negatively affect outcomes? **Task 2B** – Identify required transparency level: Select the appropriate according to the three levels of transparency (Mercado et al., 2016) based on the SAT model (Chen et al., 2014): • Level 1 no additional transparency, measures – This effectively means no requirement for transparency exists or could be identified and a more traditional design approach can be adopted.

		• Level 1+2 baseline plus descriptive reasoning and rationale or why the model is giving the advice. • Level 1+2+3 included all the previous information plus projection of uncertainty or the potential for error in the advice. **Task 2C** - Define prospective and retrospective transparency elements, sort the requirements developed into prospective and retrospective elements (Felzmann et al., 2019). Specifically, each of the requirements should be considered according to the following definitions: • Prospective transparency requirements relating to information on what input data will be used for • Retrospective transparency relating to information on why decisions were made and the origins of data **Task 2D** - Define expected responses: Given the complex and non-linear relationships between the competing design drivers in transparency integration, it is important to define how potential transparency measures might affect user performance. Define the expected impacts, relationships and co-dependencies between trust, workload, and SA in the transparency task context. **Task 2E** - Requirements definition: defining the problem in terms of specific transparency requirements.
3 Ideate	• Generate concepts to meet the transparency requirements identified in the Define stage.	**Task 3A** - Transparency solution conceptualization: Using the solution themes defined in Table 2 as a guide, match potential solutions to the transparency problem definition and requirements. The expectation for any design is that multiple solutions may be required and assessed in multiple configurations.
4 Prototype	• Develop candidate DSS solutions with integrated transparency to take forward to the test stage. As the prototypes mature, increased fidelity iterations allow operators/users more opportunity to interact with the agent, and to provide designers with more detailed feedback related to their	**Task 4A** - Prototype development: develop high, medium, or low fidelity prototypes of the transparency integration solution based on concepts from the Ideate phase. **Task 4B/5B** Test Iteration: this task calls for a fast iteration between the prototype and test stages (joint task).

	perceptions of the agent, which is the second goal for this stage.	
5 Test	• Evaluate the prototype against requirements developed in the Define stage. • Establish elements of the transparency integration that work to guide iterative design process. • Understand the potential positive and negative impact on trust, workload, and SA of the transparency measures.	**Task 5A** – Test: Define test metrics based on requirements and the projected impact of transparency measure developed in the Define stage, developing a test plan covering required metrics, and evaluating prototype DSS with enhanced transparency against defined metrics. **Task 4B/5B** Test Iteration: this task calls for a fast iteration between the prototype and test stages (joint task).

Empathy Stage Development

There are two dimensions to empathy in design – emotional and cognitive (Gasparini, 2015). Emotional empathy is perhaps closer to the traditional understanding of empathy in psychology, being an affective state derived from shared experience. Although there is some requirement for emotional empathy in the design of a product or service, Cognitive empathy, or the ability to understand the world from a different point of view, is more broadly applicable in design and the design thinking method (Gasparini, 2015). Both cognitive and emotional empathy can help the designer but, in many cases, especially in designing for complex and specialized tasks, cognitive empathy is more readily achieved. Empathy is developed through a deep understanding of user's needs and the methods of application of the product (Gestwicki & McNely, 2012). Lucas (2018) develops guidelines for developing empathy in design, especially as it relates to interview and observation. Although these guidelines are extensive and cover the specifics of empathy in design, we implement three core approach-based guidelines from this study – presence, open-mindedness, and lack of judgement.

It is important to empathize and understand the fundamentals of the task and the relevant human and machine agents in terms of transparency and time-critical decision support. Thoring and Müller (2011) define three tasks that underpin the development of empathy for the agents in a given task: interviews, observations, and interpretations. A key innovation of this framework is to encourage the designer to consider understanding of the machine agent in terms of empathy, as one would with a human user. Where we can aim to develop empathy with the human agent through interviews, observations, and development of an emotional bond, this is not possible with the machine agent. In lieu of this, the researcher needs to understand as much as possible about the specific nature of the machine agent. This involves understanding the data that drives the models, dependencies of the model on data gaps, or incorrect data, and strengths and limitations of the model. Although this may not immediately fit the understanding of empathy between humans, considering human and non-human agents in the 'empathize' stage of design thinking allows for a unified process. Though this may be somewhat of a conceptual leap for some prospective designers, we include task 1C (see Table 1) to provide guidance as they move toward a deeper understanding or 'empathy' with AI agents.

Defining the understanding of the machine agent in terms of empathy has two benefits to the design process. First, it simplifies the definition of the framework by matching the understanding of human and machine needs and dependencies, and secondly, framing this as empathizing aligns with the consideration and assessment of human and machine agents as a combined system with aligned motivations.

Define Stage Development

The aim of the define stage in traditional design thinking is to take the understanding developed in the empathy stage to define the problem and output specific requirements to guide the ideate phase. This stage aims to define the purpose of the machine agent (i.e., why the agent was developed and for what tasks will the agent be responsible), and how will the addition of an agent teammate assist operators/users? We include structured tasks to guide the designer in defining the level of transparency, type of transparency and prompt consideration of the impacts, trade-offs, relationships and dependencies between trust, workload, SA, and transparency integration. These tasks should be conducted sequentially to incrementally build a problem definition.

In the define stage, we introduce a specific task to define the level of transparency required in the design. We use the three transparency levels (Mercado et al., 2016) adapted from the three levels of the SA-based Agent Transparency (SAT) model (Chen et al., 2014). The SAT model defines the level 1 requirement for an agent to enable the human operator to understand “what is going on and what the machine agent is trying to achieve?” (Chen et al., 2014. p. 2). This relates to basic information on the operation and goals of the system, and we will consider a baseline level of transparency needed to operate and conduct operations with a machine agent. The level 2 transparency requirement is to convey rationale behind the machine agent’s decisions “Why does the agent do it”? (Chen et al., 2014. P.2). This could be information highlighting specific reasons for a decision potentially regarding task or operator characteristics or a combination of both. For instance, if the size or age of a patient was a key driver in advice given by a clinical DSS, this could be highlighted along with the specific advice. The level three transparency requirement is to give prognostic information along with uncertainty or likelihood of error, “what should the operator expect to happen” (Chen et al., 2014. p. 2). This could be providing the operator with a probability of success or a series of options to choose from combining the level two transparency requirements for each and historical outcomes in similar cases for each.

Ideate Stage Development

A key appeal of ideation philosophy in design thinking, and its application to transparency integration is the enabling of designers to develop a wide range of solutions without restriction. We use the existing design thinking ideate process to provide the approach for this stage of our framework, adopting the five core ideation philosophies from Kembel (2009): (a) share ideas, (b) all ideas worthy, (c) "yes and" thinking, (d) converge/diverge, and (e) prioritize.

This stage of the framework assists the designer in conceptualizing how the agent will operate, the level of automation-intelligibility, transparency of algorithms, and to what degree the agent will describe its reasoning. The framework maintains the flexible approach to the generation of ideas but includes several solution themes rather than a prescriptive 'toolbox' of solutions. These solution themes represent the range of approaches that may enable transparency of information in time-critical decision support but prevent rigid focus on specific implementations. For each of these themes, we establish potential advantages, disadvantages as well as requirements and dependencies. The aim of these themes is to both guide thinking and aid innovation. The themes we have defined in this framework are shown in Table 2A.2.

Table 2A.2 Transparency Solution Themes, Properties, and Potential Impacts

Transparency Solution Themes							
	Information themes	**User Interface enhancement themes**					
	Integrate additional transparency information or transfer functionality	**Enhanced provision for two-way human machine communication:**	**Remove non-critical information obscuring transparency goals**	**Modify information display to improve transparency efficiency**	**Modify the timing of presentation of information**	**Modify the display modality of information transfer**	**Other transparency integration solution**
Description	Provide explanatory information for: Algorithm transparency • Explainable AI • Uncertainty information Data transparency • Data Integrity • data Bias • Human bias	Two-way transparency requirement - enhance machine agent understanding of human agent state.	Identify information in existing systems that may obscure the intent or undermine trust between machine and human agents.	Intuitive displays to minimize workload in transparency information. Utilize users existing mental models, consistency, and recognition over recall.	In time-critical decision support, there are potentially lower workload phases that can be utilized to give additional transparency information.	Use visual, audio, and haptic displays where appropriate. The development of natural language voice assistants may enable enhanced use of the audio modality, and smart systems offer potential for haptic display integration.	To maintain the flexibility of the Design Thinking approach, we explicitly include a non-specific solution theme. This might seem unnecessary, but it ensures the designer considers solutions outside the bounds of these themes.
Dependencies	Identification of information required to improve the transparency of the specific application.	Additional modalities require displays and definitions of the information required. Input devices (manual, voice, gesture) are required.	Redundant information or inaccurate information that negatively impacts transparency, it needs to be identifiable in order to remove it.	Utilize Nielsen's heuristics of usability to make display of transparency information more efficient, effective, and interpretable.	To utilize this time, first there needs to exist lower workload times and secondly, these need to be either detected or predicted by the decision support system.	Attention resource theory suggests that separating information into multiple modalities can increase the ability of the user to attend to them.	Task/implementation dependent.
Transparency Impact	requires the attention of the user to be ultimately useful.	Enables prediction of human and machine agent state and	Reduce clutter and extraneous alerts that can distract, mislead, or contribute to alert fatigue.	More intuitive displays can make transparency information easy to interpret, reduce additional cognitive	Improve transparency by enabling attention or can undermine trust and transparency if timing is poorly	Careful consideration of the system environment and the capabilities of human agents is needed to	Task/Implementation dependent

		optimization of information. Enhances data integrity and counters operator bias		processing and reduce workload.	implemented and leads to confusion.	deconflict secondary display modalities.	
Workload Impact	Increase	Increase	Decrease	Decrease	Balance over time/minimize increase	Decrease if additional resources available	Task/implementation dependent
Trust Impact	Increase	Increase	Increase	Task/implementation dependent	Potential decrease if time delta too long	Task/implementation dependent	Task/implementation dependent
Situational Awareness Impact	Task/implementation dependent	Increase	Task/implementation dependent	Task/implementation dependent	Potential decrease	Task/implementation dependent	Task/implementation dependent
Task Performance Impact	Task/implementation dependent	Improved short term teaming/long term data integrity.	Potentially improve	Task/implementation dependent	Task/implementation dependent	Task/implementation dependent	Task/implementation dependent
Potential unintended consequences	Increase workload, reduce task performance through information overload	Contribute to information overload, decrease trust in automation	Removal of required information and change to familiar interfaces reduces trust and performance	Unfamiliar or novel displays may reduce trust in the system and fail to communicate transparency info.	Reduced trust and increased workload if timing of information is poorly predicted or too distal.	Information in unfamiliar channels may be ignored or conflict with external systems or teaming.	Task/implementation dependent.

Uncertainty and how to determine and convey uncertain information are key aspects of transparency integration. Bhatt et al. (2020) discuss the idea of introducing uncertainty into HMT as a form of transparency. This enables a broader understanding of the probability that information is to be trusted, but results in a requirement for increased cognitive resources (Bhatt et al. 2020). This leads into a second, connected problem, how to provide this information to the human agent in a way that is interpretable. Zuk & Carpendale (2007) discuss approaches to visualization of uncertainty and how it can improve performance, noting the potential hazard if such visualizations increase cognitive load.

The Ideation stage aims to conceptualize solutions for the inclusion of transparency information in time-critical tasks. Including additional information processing requires additional resources, that can impact task performance. Mercado et al., (2016) highlight an important consideration in machine agent transparency - that it should not be the goal to provide all underlying information relating to transparency, but to relate clear and efficient information succinctly to minimize the impact on workload and maximize SA in the human agent. This is particularly important in time-critical tasks as attentional resources are limited.

The requirements defined in the previous stage provide targets for the designer in the ideate stage but, the relationships and dependencies between elements can be complex and non-linear or unidirectional. The designer should be aware of this and expect some unpredictability in results. To enable users to calibrate trust in machine agents, it is important that users are aware of the machine agent's limitations and can predict its errors. Concepts should aim to provide information about the AI performance, describing what the agent does and how well it does it.

Prototype Stage Development

Traditional design thinking establishes processes for prototyping such as storyboarding, mockups, and retaining an iterative approach. Our framework maintains these prototyping philosophies to maintain the fast iteration through the test and prototype stages. There are no specific requirements incorporated into this stage of the design thinking process adaptation. Prototypes should however be developed to an appropriate level of fidelity to enable appropriate transparency information or interface design to be incorporated and tested. Prototypes will generally increase in fidelity and complexity as the iterations increase.

Test Stage Development

Our framework preserves the key elements of the design thinking philosophy to enable fast iteration prototyping. We adopt the following three tasks from the baseline design thinking model: (a) assess performance (are transparency requirements met), (b) identify what works (converge/diverge in fast iteration), and (c) understand the impact of transparency integration. Although these are the processes associated with existing design thinking process, each of the three needs to be understood in terms of trust, workload, and SA.

When designing test plans to for assessing the performance of the prototype and specifically the machine agent, it is important to consider the reliability, faults, and predictability of errors in the system to understand the impact of the transparency implementation on trust. The specifics of the test phase will be dependent on the solutions defined. We highlight the importance of understanding the impact of transparency on trust, workload, and SA. As a minimum, tests should include qualitative or quantitative assessments of all three of these constructs, along with assessments of the relationships and potential impact on Transparency.

Framework for Transparency in Time-Critical Decision Support

The framework developed in the previous section is summarized in Figure 2A.3, and Tables 1 and 2, outlining the adaptations and key tasks in a five-stage process aligned with design thinking model defined by Kembel (2009). This framework maintains the key strengths of the design thinking approach; focus on user needs, enabling innovation and fast iteration, along with additional structure to facilitate the effective integration of transparency measures.

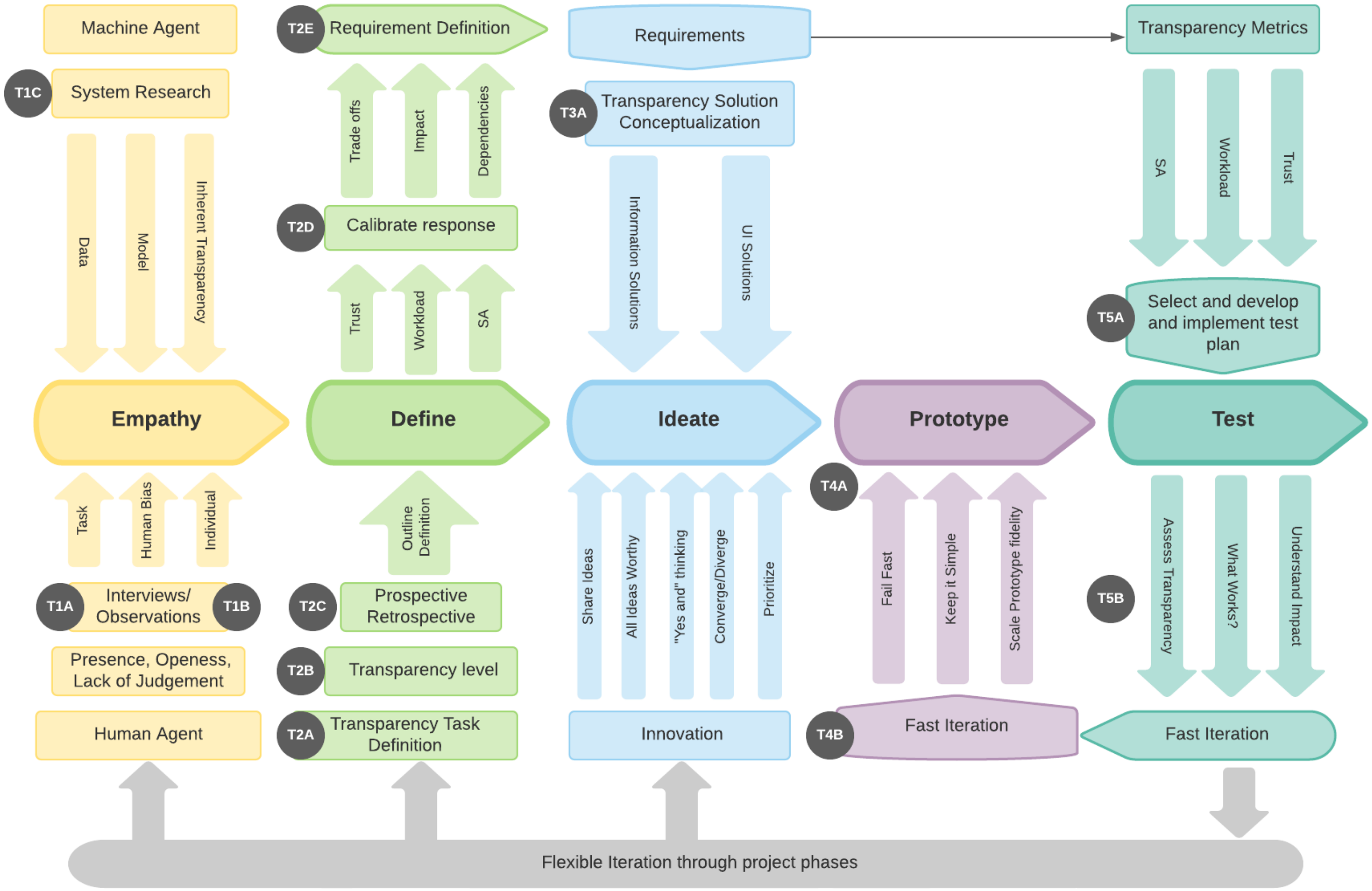


Figure 2A.3. Design Thinking XAI Integration Process – refer to Table 1 for task definitions

Framework Implementation

This implementation focuses on a concept for a digital ToC assistant, enabling Emergency Medical Service (EMS) users to record patient and injury information and predict patient states during ToC. Each stage of the implementation is discussed in detail in the following sections. This implementation consists of the first four phases of the design thinking framework, and we address the future implementation of the test phase in the discussion. The test stage was not completed due to the required involvement of human subjects in the test phase and incompatibility associated with the COVID-19 restrictions in place during Spring 2021.

Framework Implementation – Empathy Stage

Task 1A – Task Analysis

ToC is the transfer of a patient from one agent (referring) to another (receiving). Those agents can be Emergency Medical Teams (EMT), physicians in trauma centers, medical center nurses and so on. Accurate and timely transfer of information plays a key role in ensuring safe ToC. Poor transfer of a patient's information may result in poor ToC, resulting in additional healthcare cost, holding up patient recovery time, and threatening life (Karnon, 2003). The problem of transparency in ToC has several elements requiring consideration, such as identifying which aspects of the ToC decision support are lacking in transparency, identifying potential solutions to improve ToC, and analyzing the potential impact of implementing solutions to improve transparency, while considering both the impacts of solutions and the impact of the ‘do nothing’ option.

Hierarchical Task Analysis (HTA) was conducted to deepen understanding of the tasks specific to the information transfer between medical professionals during ToC. HTA was conducted through consultation with Subject Matter Experts (SMEs) based at the National Center for Medical Readiness (NCMR). Data was collected through structured interviews where, users were asked to walk through the ToC process with probes to establish detail or clarify or isolate individual task components to establish detailed understanding of the task, as well as its dependencies and user motivations. No personal information was collected during the interview process. This HTA was based on the existing, unaided ToC process to gain an understanding of the context and transparency-trust requirements. In addition, the Standard Operating Procedure (SOP) for Patient Handoff Between a Healthcare Facility and a Transporting Ambulance (2016) was analyzed to understand the operational ideals and motivators for the ToC task. The HTA is shown in Table 2A.3.

Table 2A.3. Hierarchical Task Analysis of Transfer of care task

<table>
<tr><td colspan="2"><b>0 EMS Transfer of Care to triage professional</b>
Plan 0.</td></tr>
<tr><td colspan="2">1. Assess patient—throughout emergency situation
Plan 1. Do 1.1 and 1.2—repeat 1.0, 2.0 and 3.0 as needed</td></tr>
<tr><td></td><td>1.1 Visual Inspection of Patient condition
Plan 1.1. Do 1, 2 then 3 in sequence
1.1.1 Check outward signs of consciousness, hemorrhage, breathing
1.1.2 Consult electronic heart/blood pressure indication
1.1.3 Verbal consultation with patient if able.
1.2 Recall prior condition
Plan 1.2. Do 1, then 2 in sequence
1.2.1 Recall previous patient condition from chart/record if available
1.2.2 Recall from memory if not on chart and recall is possible</td></tr>
<tr><td colspan="2">Build mental picture of patient condition—throughout emergency situation
Plan 2. Do 1, 2, if mental picture is insufficient repeat 1.1 and 1.2. Do 2.3 and 2.4.</td></tr>
<tr><td></td><td>2.1 Determine current status
2.2 Determine past status
2.3 Establish perceived condition delta
2.4 Determine potential improvement/degradation probability</td></tr>
<tr><td colspan="2">3. Perform treatment—throughout emergency situation
If 2 requires Plan 3. Do 1-2-3-4,</td></tr>
<tr><td></td><td>3.1 Establish appropriate treatment
3.2 Execute treatment
3.3 Check effectiveness of treatment
3.4 Record treatment</td></tr>
<tr><td colspan="2">4. Transfer information—on arrival at primary care facility
Plan 4. Do 1, iterate through 2-3-4, 5 for all characteristics repeat from 3 if error detected</td></tr>
<tr><td></td><td>4.1 Identify triage nurse/appropriate handoff professional
4.2 Recall patient status
4.3 Verbal transfer of individual patient characteristic
4.4 Await accurate confirmation through talkback protocol from receiving agent.
4.5 Check for error in talkback protocol
4.6 Once complete and content with accuracy and completeness of information transfer conduct formal hand over, including paperwork.</td></tr>
</table>

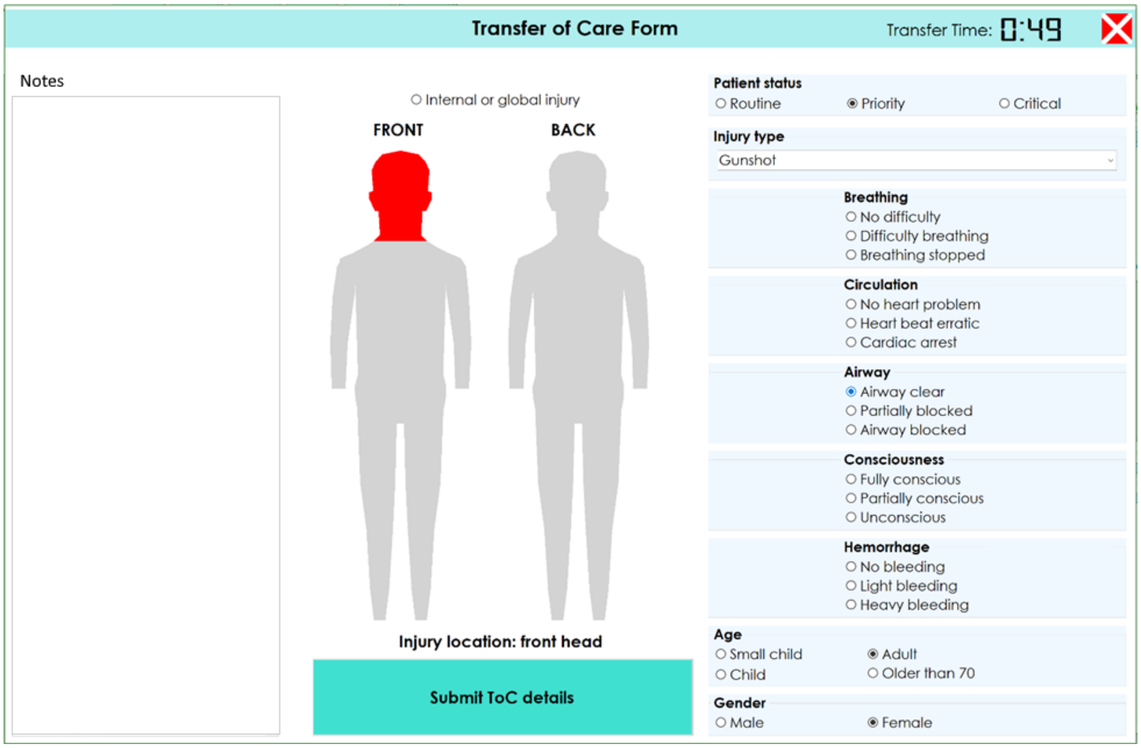
Task 1B – User Analysis

EMT, triage professionals and trainees, and military Medivac and receiving medics were considered to be the primary users of the DSS. A user analysis was conducted to further understand the characteristics, motivations, and background of the user population. Users were questioned on personal characteristics such as age, experience, visual and auditory capabilities as well as impairments. Unstructured interviews with open-ended questions were used to elicit additional detail from users on how their characteristic impact how they conduct tasks. In addition, more general data on the age ranges, gender profiles and user education levels were conducted. The main findings of this process were that clear, unambiguous information was key to accurate recall by the receiving agent. Talkback protocol, where verbal confirmation of information transfer is given by both transferring and receiving parties, is key to information assurance. The task is generally high workload but depending on the severity of the injury there are times where workload is reduced. Understanding the accuracy and integrity of the system was considered important to all users. There was some reluctance towards including personal information in a system that could be used to gather individual performance data. Visual display modalities were preferred for permanence, although auditory displays were generally considered acceptable but may suffer in louder environments. Simplicity was considered important in all aspects of a ToC DSS.

Task 1C – System Analysis

The prototype ToC DSS interface is based on a prototype previously developed (Stone, 2019). The system summarizes current patient information but also predicts potential future states based on a predictive AI model trained on data from the National Emergency Medical Service Information System (NEMSIS) database (Mann, 2016). This predictive model is the AI element requiring transparency information.

The AI in this model is primed by user input of patient parameters: age, gender, injury type (gunshot, blunt force trauma, etc.), injury location and vitals (3 level input for hemorrhage, circulation, respiratory, airway, consciousness). The model then predicts potential future patient states based on information drawn from events with correlated parameters in the NEMSIS database (Mann, 2016). The model outputs a predicted patient state at the estimated time of arrival at a primary care facility. Screenshots of the prototype ToC DSS are shown in Figures 2A.4 and 2A.5.



Transfer of Care Form

Transfer Time: 0:49

○ Internal or global injury

Injury location: front head

Patient status

○ Routine ● Priority ○ Critical

Injury type

Gunshot

Breathing

○ No difficulty
○ Difficulty breathing
○ Breathing stopped

Circulation

○ No heart problem
○ Heart beat erratic
○ Cardiac arrest

Airway

● Airway clear
○ Partially blocked
○ Airway blocked

Consciousness

○ Fully conscious
○ Partially conscious
○ Unconscious

Hemorrhage

○ No bleeding
○ Light bleeding
○ Heavy bleeding

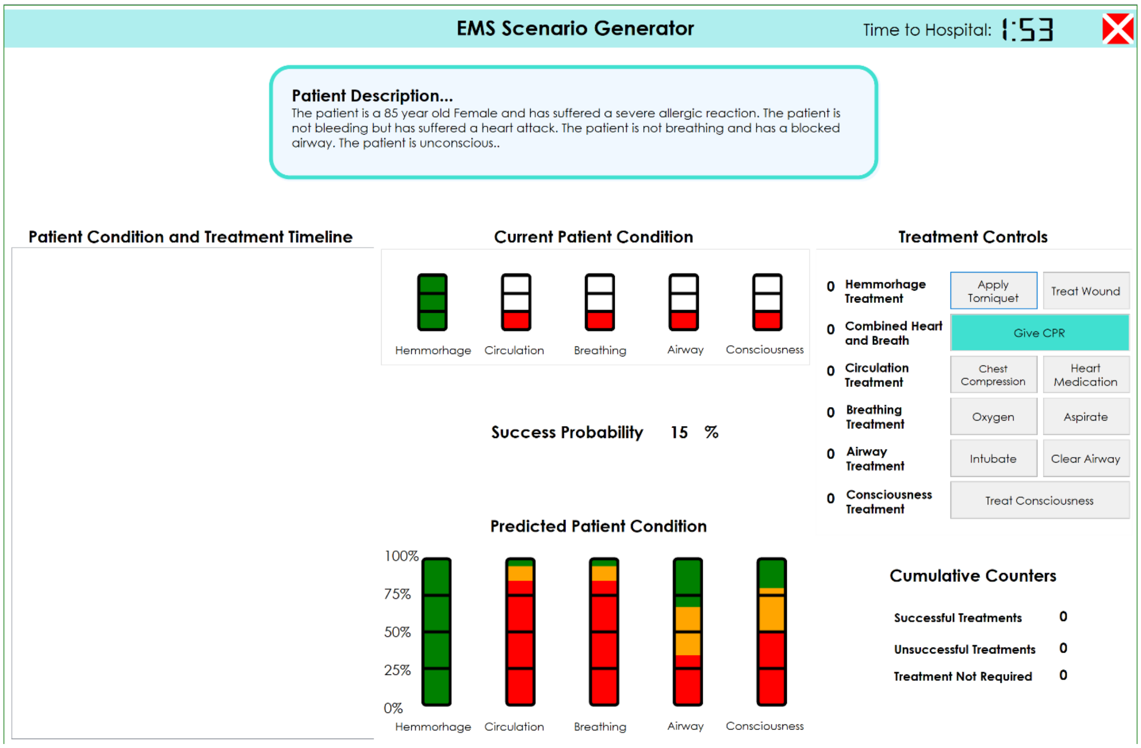
Age

○ Small child ● Adult
○ Child ○ Older than 70

Gender

○ Male ● Female

Figure 2A.4. Baseline Transfer of Care DSS – Patient Status Input Interface



Patient Description...

The patient is a 85 year old Female and has suffered a severe allergic reaction. The patient is not bleeding but has suffered a heart attack. The patient is not breathing and has a blocked airway. The patient is unconscious..

Success Probability 15 %

0	Hemmorhage Treatment
0	Combined Heart and Breath
0	Circulation Treatment
0	Breathing Treatment
0	Airway Treatment
0	Consciousness Treatment

Successful Treatments	0
Unsuccessful Treatments	0
Treatment Not Required	0

Figure 2A.5. Baseline Transfer of Care DSS – Patient Status Tracking Interface

Framework Implementation – Define Stage

Task 2A – Transparency Task Definition

What are the specific issues relating to transparency between human and machine agents in this specific context?

- The ToC task requires efficient information flow between the EMT and triage professionals.
- The task requires sensitive information to inform decision support.
- The system improves as it collects additional information on the task through reinforcement learning. Developing understanding of this goal in human agents allows for improved human-machine teaming.

What are the existing or future impacts of a lack of transparency?

- Users less likely to trust the system if predictions do not match outcomes.
- Users may be unlikely to share information with machine agents if the potential uses of the data and associated consequences are unknown.

How can these transparency implementations improve outcomes?

- Introduce uncertainty associated with prediction of patient states can allow users to calibrate their trust of the system and have a clear understanding of uncertainties in advice provided by machine agents.
- Introduction of transparency information on prospective and retrospective data use can improve data integrity and feedback for reinforcement learning as well as improve trust between human and machine agents.
- Introducing transparency measures and improving trust can enable users to better predict machine behavior, potentially reducing workload and improving SA.

How might poorly implemented transparency negatively affect outcomes?

- Poorly implemented transparency measures where users existing mental models are disrupted can negatively impact both task performance and trust.
- Poorly prioritized use of alerts to cue users to transparency information, particularly audible or haptic, can increase alert fatigue, reducing trust and increasing workload.
- ToC is a high workload task. Providing additional information can increase demand for user's cognitive resources, impact task performance and undermine trust.

Task 2B – Identify Required Transparency Level

The transparency task definition highlights the need for both descriptive transparency measures and uncertainty information so in this case we define the required transparency level as Level 3.

Task 2C – Define Prospective and Retrospective Transparency Elements

The prospective transparency elements are comprised of use of user information for performance assessment and use of task input information for ToC task. The retrospective transparency elements are comprised of uncertainty in vital state predictions provided by the DSS, biases in existing data, incomplete or incorrect data, and sparse data.

Task 2D – Calibrate Responses

Expected relationships between constructs:

- Improving transparency improves trust in the system (Mercado et. al. 2016; de Fine Licht & de Fine Licht, 2020).
- Improving transparency improves user performance (Mercado et. al. 2016).
- Additional transparency information increases workload and reduce the availability of cognitive resources (Helldin, 2014).
- Increased SA improves task performance (Ikuma et al., 2014).
- Increased trust improves SA (Yoko, 2006).
- Increased workload can reduce SA in the time critical task and vice versa (Endsley, 1995).
- Expected impacts derived from these relationships:
- Improved trust in ToC DSS
- Improved SA in both EMT and triage professionals
- No effect on task workload

Task 2E – Requirement Definition

Requirements:

- The system will provide uncertainty information to the user to improve trust calibration in the vital state prediction display.
- The system will provide information or alerts to the user to notify sparse or incomplete data driving advice provided by the system.
- The system will inform the user of the limitations of use of required user information and inform of all legal information.

- The system will prompt the user if required input task information (user actioned) is missing.
- The system will alert the user to potential biases in existing data (e.g., age or location related variations in decision support advice).
- The system will provide additional transparency advice without increasing task workload.

Framework Implementation – Ideate Stage

Task 3A – Transparency Solution Conceptualization

Requirement 1: The system will provide uncertainty information to the user to improve trust calibration in the vital state prediction display.

Solution concept:

- Integrate additional transparency information: The existing DSS provides projected patient status probabilities for each vital statistic indicating the percentage likelihood that the end state of the patient will be normal (green), poor (orange) or critical (red). Introducing an error bar based on one standard deviation of the underlying data will give the user an indication of how uncertain each parameter projection is.
- In addition to this, the DSS provides an estimated recovery probability. Here a less complex transparency implementation detailing a recovery probability range, utilizing a single standard deviation could provide uncertainty information with minimal additional attentional requirement.

Requirement 2: The system will provide information or alerts to the user to notify sparse or incomplete data driving advice provided by the system.

Solution concept:

- Integrate additional transparency information: Introduce a red, orange, green scale labelled ‘Data Integrity’ to highlight potential issues (orange) and definite issues (red). When the indicator is green, no action is needed. Implementing an interrogatable warning where the indicator can be clicked to provide additional information could enhance transparency and minimize the increased workload, allowing the user time to determine the detail of the problem.

Requirement 3: The system will inform the user of the limitations of use of required user information and inform of all legal information.

Solution concept:

- Integrate additional transparency information: Provide a clickable information button with a short cue phrase to highlight use of data and provide the means to interrogate the interface further. No alerts will be associated with this, and it is assumed users will utilize this transparency functionality if needed.

Requirement 4: The system will prompt the user if required input task information (user actioned) is missing and highlight the specific areas requiring attention.

Solution Concept:

- Enhanced provision for two-way human machine communication: The system will provide additional input mechanisms in the interface to enable users to input key data more easily. Initially, this will be a passive visual display option. Where vital information regarding patient status is missing, the system will include provision for a voice alert combined with either voice or manual input modalities. Only one audio alert will be given but a visual marker indicating the missing information will be provided. This will include a secondary transparency message to inform the user of the impact of the missing information.

Requirement 5: The system will alert the user to potential biases in existing data (e.g., age or location related variations in decision support advice.)

Solution Concept:

- Integrate additional transparency information: Introduce a red, orange, green scale labelled ‘Data Bias’ to highlight potential issues (orange) and definite issues (red). When the indicator is green, the user does not need to take action. Implementing an interrogatable warning where the indicator can be clicked to provide additional information could enhance and minimize the increase workload, allowing the user time to determine the detail of the problem.

Requirement 6: The system will provide transparency advice without increasing workload.

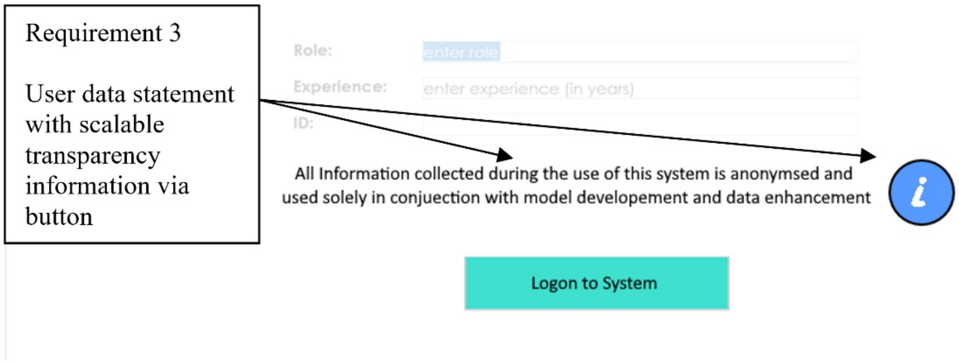
Solution Concept:

- All transparency solutions are designed to minimize time dependency and additional cognitive resource requirement. Where provided, additional transparency information is scalable – immediate alert of potential issues with further detail available when user has cognitive resources to spare.

Framework Implementation – Prototype Stage

Task 4A – Prototype

A low-fidelity prototype was developed in photoshop to include the transparency concepts outlined in the ideate stage, specifically the input and patient vital displays. The transparency measures included in the prototype including transparency measures are annotated in Figures 2A.6 and 2A.7 (baseline interface elements are shown greyed out for clarity).



Requirement 3

User data statement with scalable transparency information via button

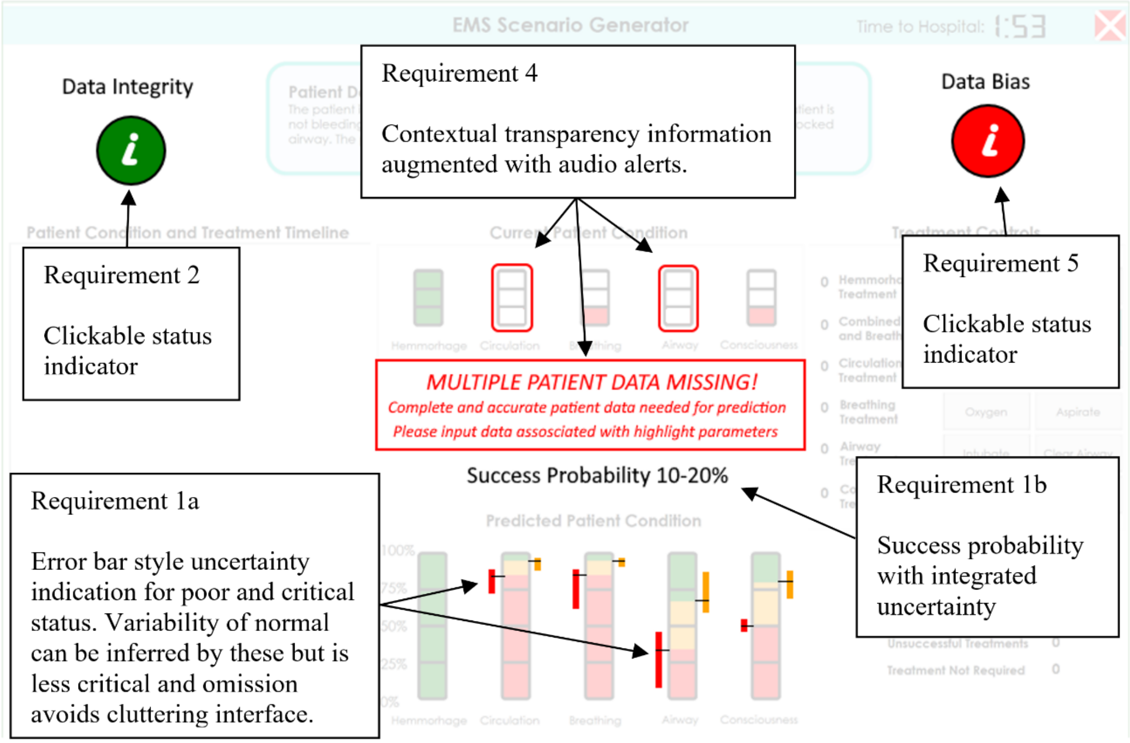
Role: enter role

Experience: enter experience (in years)

ID:

All Information collected during the use of this system is anonymsed and used solely in conjuection with model developement and data enhancement

Figure 2A.6. Transfer of Care DSS with Transparency Measures – Patient Status Input



Requirement 4

Contextual transparency information augmented with audio alerts.

Requirement 2

Clickable status indicator

Requirement 5

Clickable status indicator

Requirement 1a

Error bar style uncertainty indication for poor and critical status. Variability of normal can be inferred by these but is less critical and omission avoids cluttering interface.

Requirement 1b

Success probability with integrated uncertainty

MULTIPLE PATIENT DATA MISSING!
Complete and accurate patient data needed for prediction
Please input data assosciated with highlight parameters

Figure 2A.7. Transfer of Care DSS with Transparency Measures – Patient Status Tracking

Discussion

Transparency in AI-based systems is essential for effective HMT. This is especially true in scenarios that are inherently risky and time-critical (e.g., ToC during emergency medical treatment). It is imperative that the information users receive from the systems they work with be interpretable, and that contributing factors and rational that lead to the agent’s recommendations and/or actions are clear. We have proposed a framework for the integration of transparency measures in time-critical decision making, which takes into account the complex ways in which transparency influences trust, workload, and SA, and demonstrated an implementation of the first four stages of this framework in a prototype ToC DSS. The framework maintains the advantages of design thinking and integrates additional structure to streamline design choices. We establish three novel adaptations to design thinking to accommodate transparency integration. First, we consider all agents, both human and machine as equal stakeholders when developing empathy and a deep understanding of the problem. Second, we define key constructs for design and test of transparency in time-critical decision support: trust, workload, and SA.

In addition to these measurement constructs, we include definitions of three transparency levels and the concepts of prospective and retrospective transparency to structure our adapted design thinking. We establish the complex, non-linear and task dependent nature of these relationships and rather than prescriptive design rules, suggest these be considered as trade-offs in the design process. Finally, we establish solution themes which guide the designer in thinking about the methods, both to integrate transparency, and mitigate impacts across the key constructs we have defined.

As AI becomes more powerful and complex, the need for transparency designers to develop a deep understanding of the model, its purpose, and the underlying data becomes even more important. Including machine agents in the empathy stage of design thinking aims to ensure that consideration of both humans and machines in the design stage. To ensure AI is human-centric, the transparency designer must ensure that this understanding is passed in suitable detail to the end-user.

We believe that this framework can be followed to expedite the integration of transparency in time-critical decision support. Allowing flexibility and adaptability to produce task specific solutions but giving enough structure and understanding of the potential benefits and drawbacks of solutions to enable transparency requirements and solutions to be more quickly developed and tested. Although we have established evidence for the contributions to this framework, and demonstrated the first four stages, it remains to demonstrate the testing stage to understand the entire framework. We therefore propose a case study implementation of the framework test stage.

As noted, the test phase was not completed due to the impossibility of human testing due to COVID-19 restrictions in place during Spring 2021. The elements of the implementation task conducted used task analysis and user profiling in the empathy stage from the baseline system development. As the tasks are identical, we believe this is not a major limitation. The remaining stages formed most of one iteration of the design thinking framework for the integration of AI transparency and were completed in 2 days. In a full iteration, the empathy and test stages are likely to be more time consuming and complex than these stages. During the implementation of the framework, it was found that the define stage needs less restrictive, definition of levels and types of transparency as a system might require different levels of transparency across all individual requirements.

Once human subjects testing becomes possible again, we aim to conduct user testing on the prototype ToC decision support system, developed in the current study against the baseline prototype (Stone 2019). We believe this will provide evidence to support our assertion that trust, workload, and SA are key to successful integration of transparency measures in time-critical decision support. Establishing time-criticality in the future study is key to demonstrating this framework in the context for which it is designed. It will therefore be important, in so far as is possible, to represent the environment and pressures of the real-world task and conduct assessments with participants drawn from the medical professional community with experience in ToC.

Conclusion

We successfully developed and implemented a framework for the design of transparency measures in an AI-based DSS based on the design thinking framework. Only one iteration of the prototype and test was possible and will require further testing to determine how successful the framework is in the fast iteration phase. Overall, the transparency integration framework has shown the potential to ensure robust integration of transparency measures in time-critical AI-based decision support.

Please see the Reference section at the end of the document

End of Paper: Development of a Novel Hybrid Cognitive Model Validation Framework for Implementation Under COVID-19 Restrictions

Chapter 3 – A Complete Design Framework for XAI in TCS

The first two chapters, and the associated papers establish that XAI is a complex problem, and one that the research community is just beginning to address. While this means there are opportunities for original research, the vastness of the unknown is somewhat daunting. In the previous paper we established a design framework for XAI, and this chapter extends that framework by instantiating the framework in a real-world problem and validating the tools and protocols defined in the framework through a user study.

To determine where to focus this stage of the research, I looked back at the XAI design framework in paper three, and noted three of the five design thinking stages that would most obviously benefit from additional development – Empathy, Prototype and Test, which, as of the end of chapter 2, did not have XAI specific design elements, see Figure 3.1. This is more an incremental focus than the evolution between the background development of chapter one, and the crystallization of the research concept in chapter 2, but still addresses a key development area of the XAI framework.

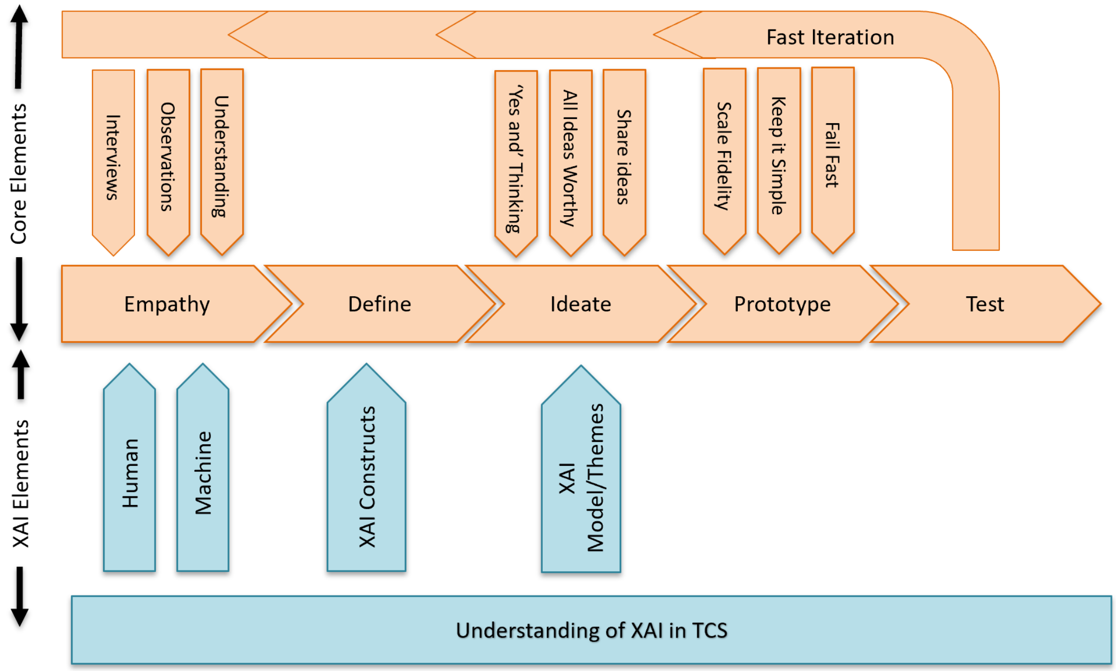


Figure 3.1 Overview of the Design Thinking Framework at the end of stage 2.

I felt at this stage that the most important problem to consider was the development of a means to assess the performance of XAI in TCS, and decided to focus chapter 3 on the development of the prototype and test stage of the XAI design framework.

In this chapter, therefore, I address the need for assessment methods compatible with the fast iteration requirement and TCS in general. Specifically, I focus on measures and metrics for the underlying constructs of XAI in TCS, specifically trust and workload. This updated version of the XAI design framework provides a structured approach to the design of XAI systems, with bespoke tools at each of the design phases tailored to the problem of integrating XAI in TCS. The additional elements of the XAI design framework covered in chapter 3 are shown in Figure 3.2.

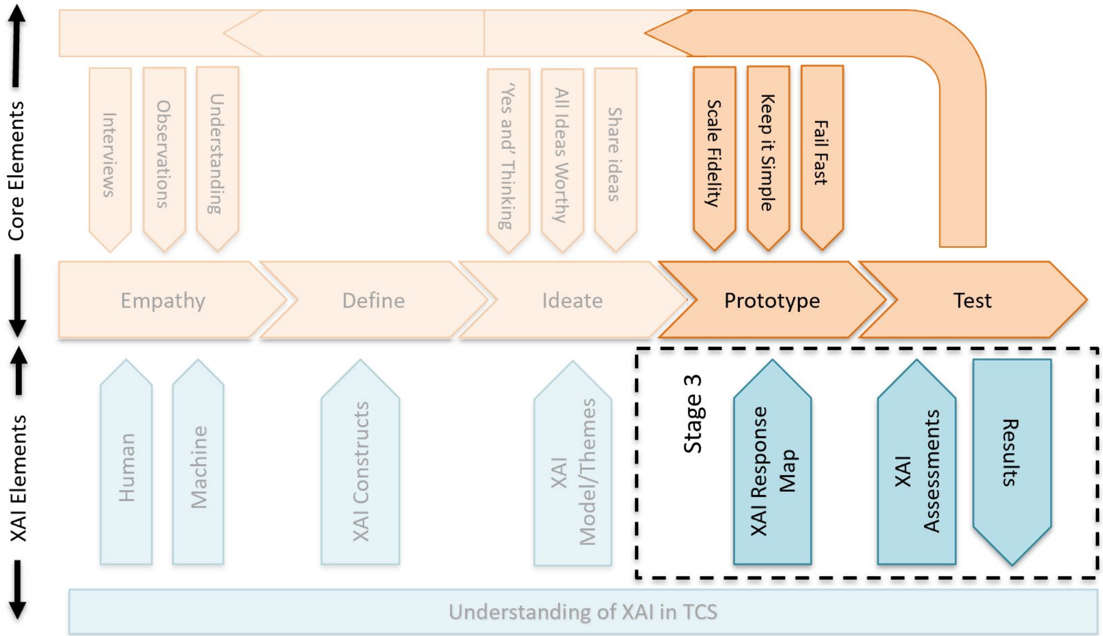


Figure 3.2 Stage 3 XAI Design Framework concept

The paper presented in this chapter details the development of these assessment methods. This covers the conceptualization of assessment methods, through to the development of an experimental program to provide evidence for their effectiveness and determine if they should be included in the XAI design framework.

In the context of the AI copilot example, this chapter aims to provide the designer with a means to determine whether XAI measures included by the designer increase task performance, along with an assessment of the potential impacts on trust and workload, which might have positive or negative implications for the drone pilot task. Using physiological measures is particularly important for tasks such as drone pilotage, and TCS in general where subjective, questionnaire based methods are difficult to implement without affecting the nature of the task. Including these elements in the framework is key to a fast iterative design framework, enabling designers to make informed, evidence-based updates to prototypes.

Eye Tracking for Explainable AI (XAI) performance Assessment in Time-Critical Systems.

by

Paul Stone

Wright State University

2022

Wright State University

ABSTRACT

Artificial Intelligence (AI) has seen a surge in popularity as increased computing power has made it more viable and useful. As AI becomes more complex, the rationale behind decisions and the relationship between the data and model output can be harder to interpret. This can result in incorrect predictions, classifications, or interpretations, leading to over-reliance, under-reliance, or confusion. Additionally, AI models can contain algorithmic, data and design bias, which may exacerbate negative outcomes, particularly minority populations.

Explainable AI (XAI) aims to mitigate these problems by providing information on the intent, performance, and reasoning process of the AI. Where time is limited, or cognitive resources are highly utilized, additional information can negatively impact performance. Ensuring XAI information is intuitive and relevant allows the user to quickly calibrate their trust in the AI, in turn improving trust and task performance, and reducing workload.

This study details a human-subjects experiment do establish physiological assessment metrics for XAI in time-critical systems. More specifically, gaze entropy and gaze duration are considered as candidate metrics for workload and trust, respectively. These metrics are compared to the performance of baseline subjective measures of trust and workload to establish evidence for their use in an XAI design framework.

Introduction

Artificial Intelligence (AI) has seen a surge in popularity in the last decade since increased computing power made it more viable and useful. In addition to this, the influence and impact of AI is growing day-by-day, and the importance of explainability in AI begun to gain traction in both regulatory (EU, 2018; Kratsios, 2018) and research organizations (Ezer et al., 2019; Gunning & Aha, 2019). Although there are benefits associated with AI, there are also potential drawbacks. As the power and complexity of AI develops, the rationale behind the decisions and the relationships between the underlying data and model output can be harder for human users to interpret, (Strobel, 2019). Complex AI models can contain unseen algorithmic bias and lead to over-reliance (Kim et al., 2020) and automation surprise (Parasuraman & Riley, 1997; Sarter, et al., 1997). Despite the power to analyze large data and form complex representations and predictions, there is often little clarity provided to users about how the output is determined (Topol, 2019). This can result in incorrect or harmful predictions, classifications, or interpretations by AI models (Danks and London, 2017; Garcia, 2016) and an inability for users to identify these failures (Kim et al., 2020; Parasuraman and Riley, 1997). Even where there are no specific failures, decision makers demand explainable systems (Hoffman et al., 2018), while users have a right to know why decisions that affect them are made (Goodman and Flaxman 2016; European Union 2018).

Explainable AI (XAI) is a developing field of research that focuses on ensuring that an AI interface provides transparency to the user through interpretable information on the rationale of the model along with its output (Doran et al., 2017). This information allows the user to determine whether to trust AI information (Xu et al., 2019). However, providing additional information requires additional cognitive resources and has the potential to cause unintended negative consequences due to increased workload (Ha et al., 2006; Helldin, 2014; Parasuraman & Riley, 1997). The potential for negative impacts of increased information is amplified in high workload or time-critical tasks, but the need for XAI and model transparency is not reduced (Wachter et al., 2017). The potential for XAI to provide improvement in human-AI teaming is highlighted by Mercado et al. (2016), who demonstrate a significant multivariate improvement in operator performance through the implementation of transparency in AI. Currently, there is debate as to the meaning of XAI and transparency, with some using the terms interchangeably and some highlighting significant differences between them. Mohseni, Zarei & Ragan (2021) define XAI as a potential solution to the need for accountability in AI, through the provision of interpretable information on AI decision-making processes and logic to system operators, noting the difficulty this may present given the diversity of tasks to which XAI might be applied. Adadi, & Berrada, (2018) more broadly define XAI research as a means to improve trust and transparency in AI-based systems and highlight the importance of explainability in the continued progress of AI. Throughout this research, transparency is considered as a means to achieve XAI, particularly at the interface level. To ensure clarity and consistency throughout the research, the following definitions were adopted:

> *"Explainable AI is a method for improving trust and transparency on AI systems, to improve accountability and decision-making performance of Human-Machine teams."*
>
> *"Transparency is the quality of an interface pertaining to its abilities to afford an operator's comprehension about an intelligent agent's intent, performance, future plans, and reasoning process" (Chen et al., 2014, p. 2).*

Transparency offers a means to ensure systems remain human-centric and empower operators and is a key aspect of building trust in both human-human and human-AI teams (de Fine Licht & de Fine Licht, 2020; Mercado et al., 2016). Achieving transparency in information systems is important in preventing negative outcomes and creating accountable systems (de Fine Licht & de Fine Licht, 2020).

A design framework for the integration of XAI information into TCS (Stone et al., 2022) was developed to provide a structured approach to XAI integration. The output framework consisted of a five-phase approach based on design thinking (Kembel, 2009) and focusing on transparency as key to achieving XAI. The initial iteration of the XAI design framework focused on the first three phases, providing a structured approach to the definition of the design problem and associated requirements through to a structured approach to solution conceptualization. The final two stages of the original design thinking approach (Kembel, 2009) on which the XAI framework was based, called for a fast-iterative approach to prototyping and test, no additional XAI specific additions were for these stages.

The main contribution of the XAI design framework (Stone et al., 2022) was model of time-critical XAI systems (TEXAS). This model has three XAI levels based on the two-level SA-based Agent Transparency (SAT) model (Chen et al., 2014), but augmented by a third XAI level combining the first two. The TEXAS model levels are:

- Level 0 - no XAI information.
- Level 1 - XAI information, or 'Rationale' information enables understanding of "what is going on and what the machine agent is trying to achieve?" and "Why does the agent do it?" (Chen et al., 2014. P.2).
- Level 2 XAI information, or 'Certainty' gives certainty or likelihood of error, and answers the question "what should the operator expect to happen?" (Chen et al., 2014. p. 2).
- Level 3 is a combination of both Level 1 Rationale and Level 2 Certainty.

The specific interpretation of these levels is task specific, but this model helps to conceptualize how XAI and transparency information can be scalable and appropriate to the demands of the specific task.

Explainable AI goes beyond these simple level classifications and is itself a complex, multifaceted problem (Helldin, 2014) covering problems as diverse as data bias (Datta et al., 2015; Johnson, 2020), explainable algorithms in AI (Kim et al., 2020), ethics and regulation (EU, 2018; Kratsios, 2018; Larsson & Heintz, 2020) and the open sharing and reproducibility of results (Haibe-Kains, et al., 2020). Designing human-centric XAI requires an understanding of constructs such as trust, workload, situational awareness (SA) and task performance. Each of these problems is complex, interrelated and can vary from task-to-task, especially in TCS (Wachter et al., 2017). While there are numerous studies looking at the need for XAI and the development of technical solutions such as SHAP (Lundberg & Lee, 2017). and GRAD CAM (Selvaraju, et al., 2017), there is a notable gap in the understanding of the implementation and implications of technical solutions.

That is, which solutions are appropriate for a given task and how and when to implement them when designing user interfaces.

Many approaches to XAI focus on the visualization of analytics such as data labelling (Bernard, Hutter, Sedlmair, Zeppelzauer, & Munzner, 2021), visualization of the classification process through saliency maps (Mundhenk, Chen & Friedland, 2019). In the case of image classification, the goal of explainability is to provide information relevant to both the image and the class of interest. (Doran, et al., 2017). The key requirement of XAI is to provide additional information to the user about how the goals of the system, such that they exist, were derived and why (Chen et al., 2014). This requires an understanding of exactly how the underlying model was built, but here there needs to be care taken to ensure that the information presented in the system is readily interpretable by the user (Zuk & Carpendale, 2007).

The problem of XAI is never more important than when designing interfaces for Deep Learning models, which are often seen as "black-box" systems (Adadi & Berrada, 2018; Kim et al., 2020). It also follows that explaining the rationale behind the model and its decisions to the user becomes more difficult as the complexity of the underlying model increases (Zuk & Carpendale, 2007). In many AI systems, the performance of the system can be defined in terms of the model accuracy on test data. Providing uncertainty information to the user, to establish a level of transparency and XAI gives an idea of the performance but needs additional context related to the reasoning process (Bhatt et al., 2020). It is also important to ensure that any transparency information is interpretable as end-user understanding of XAI and its underlying constructs might be highly variable.

This increases the demand on cognitive resources and hence workload (Bhatt et al. 2020). Time critical systems were chosen as the focus of this design framework as they represent the most demanding implementation of XAI and transparency, where users may not have time or resource to attend to additional information. Not only is there the impact of XAI information on trust and task performance, but there is also the impact of the introduction of additional information into a task context where the human agent is subject to high temporal demand.

Time-critical tasks increase the need for human operators to rapidly process information and make decisions (Gusenleitner, et al., 2019), potentially resulting in errors and delays (Horvitz & Barry, 2013) and impacting the legitimacy of information (de Fine Licht, 2011; de Fine Licht & de Fine Licht 2020), which can in turn affect trust. In experts, time-critical decisions are often characterized by recognition-primed decision making (Klein, 1993) but is more generally classified by a high temporal demand and a requirement to complete a task in a finite time (Teh, et al., 2014). As temporal demand increases, the availability of cognitive resources diminishes (Rivero, 2014; Wachter 2017), potentially resulting in automation bias and over-reliance (Cummings 2017; Goddard et al., 2012).

Trust, workload, and SA were highlighted in the development of the XAI design framework as key constructs in the assessment of XAI integration. Transparency is key to XAI and an antecedent to trust in both human-human and human-machine teams. Trust is a social construct that has been identified as important for not just the development of interpersonal relationships (Jones & George, 1998; Lewicki et al., 2006; Simpson, 2007), but has also been identified as important for interactions with non-human entities such as machines (Muir, 1987), robots (Hancock et al., 2011), automation (Schaefer et al., 2016), and AI (Glikson & Woolley, 2020). Trust is defined as a willingness to be vulnerable to another, without the capability to monitor their actions (Mayer et al., 1995). Trust is particularly necessary when there is an element of risk and the trustor (the one who is trusting) needs to interact with another party or entity (the trustee) to accomplish a task or foster a relationship for long-term interactions (Mayer et al., 1995). Existing measures of human trust in automation (Jian, Bisantz & Drury, 2000) rely on subjective questionnaires, delivered after the task. This is not ideal for the time-critical task as it requires the task to be interrupted and does not give real-time variations in trust, which is important when it comes to measuring the trust calibration of specific XAI information. As such, this research aims to demonstrate a physiological measure of trust in automation in TCS that can provide real-time variation in trust metrics.

In considering trust, and in particular human trust in automation, it is necessary to consider the differences between human-human and human-agent trust. Lewandowsky et al. (2000) and Madhavan & Wiegmann (2007), found differences in human-machine trust regarding formation, violations, and repairs, noting that that humans instill more trust in machine agents during initial interactions than they do during human-human interactions. In otherwards, people have a higher baseline of trust in machine agents when compared to people. A second way trust differs between humans and agents is in when there is a trust violation. If trust is ever violated, people tend to lose trust quicker in a machine and it takes longer to re-establish, compared to a trust violation from a human. People are generally more forgiving of other humans than they are of machine agents (Lewandowsky et al. 2000).

As humans become more familiar with either a human or machine providing information, they dynamically calibrate their level of trust in the system, in both the provider, and the information itself (Wachter et al., 2017). Trust calibration specifically refers to “the correspondence between a person’s trust in the automation and the automation’s capabilities [or performance]” (Lee & See, 2004, p. 55). Over-trust, can lead to high compliance to potentially false information, also called misuse where under-trust, or disuse, can lead to failure to follow good advice or inefficiency in the HMT (Parasuraman & Riley, 1997). Properly calibrated trust allows for effective HMTs (Parasuraman et al., 2000; Sheridan & Verplank, 1978), maximizing the potential of the automation to lower workload for operators and improving the HMT performance (Balfe et al., 2015; Lewis et al., 2018). Explainable AI, and the provision of transparency information aims to enable faster trust calibration, even down to individual decisions or single information points. Measurement of trust and task performance are key to the assessment of trust calibration and therefore critical in determining successful implementation of XAI.

Workload is also an important construct to consider, as it is key to determining the impact on the user of the additional XAI information in TCS. This is important as time-critical tasks increase the need for human operators to rapidly process information and make decisions (Horvitz & Barry, 2013), which can in turn affect trust. Similar to trust, there is a need for a physiological measure that is compatible with the assessment of workload in real-time, without reliance on post-experiment questionnaires.

Situational Awareness is also an important construct identified by Stone et al. (2022), however no physiological measures for SA were identified in the literature. As such, this research focusses on the development of evidence-based physiological measures for trust and workload, compatible with the assessment of XAI in TCS.

Significance of the research

XAI offers a means to ensure automation remains human-centric. Through XAI, the benefits of automation can be maximized while enabling human operators to mitigate the potential unforeseen negative impacts that might arise, and more accurately and quickly calibrate their trust. (de Fine Licht & de Fine Licht, 2020; Lee & See, 2004; Mercado et al., 2016). Trust is a key antecedent of XAI and transparency, but the impact of workload is also important, particularly in TCS, where there is potential for XAI information to negatively impact task performance. As a result, the need for XAI needs to be balanced with the impact on a human operator's cognitive resources and the resulting impact on trust and workload. Therefore, to ensure XAI is beneficial in the context of TCS, it is important that robust, evidence-based assessment measures and metrics are available to inform the development process.

Provision of evidence based XAI assessment metrics allows researchers to determine the impact of designs on trust and workload simultaneously and in real-time, using objective, physiological measures has potential benefit in all human-machine teaming assessments, not just XAI or TCS. More broadly, providing engineers and designers with an enhanced understanding of the implications of integration, along with tools to guide the process can ensure effective human-machine teaming without compromising task performance.

Integration of these assessment techniques into the XAI design framework (Stone et al., 2022) provides XAI specific tools at each of the 5 design thinking stages and ensures the framework can guide XAI design throughout the process. This solution feeds directly into the design XAI to counter the problems associated with the implementation of complex AI systems such as algorithmic bias (Garcia, 2016), over-reliance (Kim et al., 2020) and automation Surprise (Parasuraman & Riley, 1997). The significance of this contribution is to ensure that XAI does not come at the cost of task performance and that unintended consequences can be better predicted and avoided. Providing an easy-to-use framework to guide designers in the implementation of XAI, allows those who are less familiar with the specific problems associated to quickly define the XAI elements of their particular problem and guide them through the concept ideation and prototype development and test phases. This may be particularly important for user experience designers or AI developers in bridging the gap to XAI and ensuring human-centric design. This understanding has implications for the integration of AI into almost any decision support or supervisory control application. The ability to know when to trust automation, and quickly and accurately calibrate trust in specific information, could be of benefit to all decision makers and supervisory control operators.

Aims

The main aim of this study is to establish evidence-based physiological measures for trust and workload to be used in the assessment of XAI performance in TCS. Secondly, the study aims to test the performance of an instantiated XAI system developed by the design thinking framework (Stone et al., 2022) and understand the impact of XAI and transparency measures in TCS. System reliability is the key antecedent to trust considered in this study and it is expected that these are positively correlated, i.e. trust increases with increased system reliability. Similarly, workload is expected to be positively correlated to its antecedent temporal demand. The final aim of this study is therefore to establish these relationships between task performance, trust, and workload, by varying reliability and temporal demand, to provide better prediction of system behavior and guide prototype iteration. To meet these aims, the following specific objectives and research questions were identified:

- Identify or develop candidate objective physiological measures for trust and workload.
- Determine if the proposed physiological measures are predictors of trust in TCS.
- Determine if the proposed physiological measures are predictors of workload in TCS.

Understand the relationships between XAI, human trust in automation, workload, and task performance in TCS including the impact of XAI.

- Map the response of workload, trust and task performance and create an indicative model of XAI task performance.

In addition to these core objectives, this study will examine the performance of an instantiation of XAI information developed using the XAI design framework (Stone et al., 2022), specifically to determine the effect on task performance, trust, and workload in TCS, along with the usability of the XAI information presentation.

Physiological Measures.

Workload and trust were previously identified as key performance measures for XAI in TCS (Stone et al., 2022). Subjective, questionnaire-based methods such as the NASA TLX (Hart & Staveland, 1988), see Appendix A, and the Adapted trust in automation scale (Jian, Bisantz & Drury, 2000), see Appendix B are for measuring workload and trust respectively, and have been previously validated. These methods are used in this experiment to baseline workload and trust against the physiological methods required to assess operational performance in TCS. The subjective questionnaire methods are not suitable for performance assessment of TCS as they are intrusive, requiring participants to complete questionnaires, either during or after tasks.

A less disruptive, and potentially more objective method of trust in automation assessment using eye-tracking was proposed by Lu and Sarter (2019). In this study, eye-tracking and specifically gaze times were shown to predict trust in automation. Similarly, eye-tracker measurements in the form of gaze entropy and pupil diameter (Wu et al., 2020) are the presumptive physiological measures of workload. The authors note that gaze entropy increased with workload, with a correlation factor of 0.51.

Eye-tracking is not the only physiological measure that could be used to measure trust and workload in human-AI teams, but it offers a solution that requires low preparation times, and less intrusive set-up for participants. The assessment methods proposed for trust (Lu and Sarter, 2020) and workload (Wu, et al., 2020), offers the potential for a combined XAI performance measurement in one system. Finally, a suitable eye-tracker was available in the human performance laboratory at Wright State University. No alternative physiological measurement systems, such as EEG, or heart rate were considered at this stage of the study.

Assessment Platform Development

To enable the assessment of XAI information in terms of trust, workload, and task performance a suitable assessment platform was needed. Existing approaches to similar problems were considered, however given the complexities of the XAI task and the requirement to conduct the study using non-expert participants recruited from the university population, a generic assessment platform, with a simple, time-critical system and AI-assisted decision support elements was developed. The platform's key requirements were for participants to be required to make a time-critical decision, with the provision of AI advice that they could choose to trust or ignore. Trusting, the AI should provide a time advantage if correct and trusted by the participant.

This approach requires careful design of the temporal demand and task complexity of the game and the associated decisions, along with how the AI can assist the decisions both with and without XAI information. Finally, ensuring that non-expert participants are suitably familiar with the type of decision to ensure the trust calibration can be achieved reliably with minimal training is important. A literature review failed to identify any suitable existing trust or AI-decision support assessment games that meet these requirements. This was due to the combination of trust modulation and time-critical decision requirements. There are many examples of games that aim to assess trust between human agents or trust between humans and machines and/or automation and games that assess human decision making. These are classified as trust games and tactical decision games respectively and they both offer a means to assess elements of the impact of XAI integration, but neither is a complete solution.

Trust Games

Trust games have been widely used in behavioral economics and psychology. A well-known example is the prisoner dilemma (Poundstone, 1993) , which is a thought experiment that proposes a situation where two participants adopt the role of the trustor and trustee in an experiment. If the parties collaborate, both can improve their outcomes, however if either or both break the trust, the total payout is lower, but one party can still increase their success. This has had many different variations, but the key is that there is a potential negative outcome, and the participants have a level of vulnerability in the game – that is that the user must have a risk-reward trade-off within the game.

Tactical Decision Games

One approach to the problem of development of a suitable assessment platform for Transparency in AI decision support in high workload environments is a Tactical Decision game tailored to the requirements of the problem and the associated transparency levels. Tactical Decision games are widely used in research to train and assess decision makers and the performance of decision support systems. There is inherently an element of trust in decision making when a user acts on advice given either by a human or machine, however there was no specific instances of research that attempted to modulate or assess trust as a variable in tactical decision games. Crichton et al., (2000) demonstrated the use of tactical decision games to train staff working in high reliability industries in decision making, while Hinrichs, et al. (2021) considered the implementation of machine learning in tactical decision games, noting specific difficulties in the application due to the structure of the games. These types of games allow the decision choices of participants to be assessed in a controlled environment and the potential to modulate transparency information in such a game is applicable to this problem.

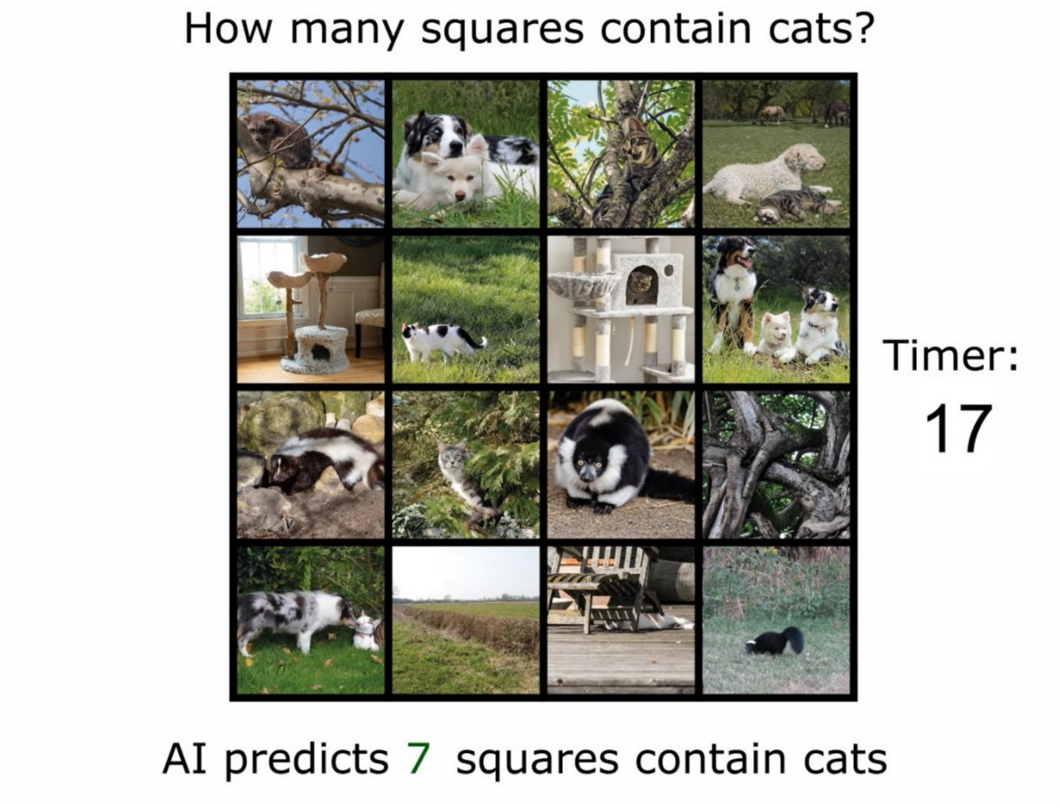
Hybrid Tactical Trust Decision game

To build a game suitable for the assessment of XAI in TCS, elements from both trust and tactical decision games were combined. In this case, the human agent and AI agent must collaborate in a supervisory control role, where the human is the trustor with the AI providing information being the trustee. The game will require a user to make decisions based on advice from an AI but will contain the elements of risk-reward characteristic of trust games. The following requirements were developed to ensure the assessment platform enables the assessment of trust, workload, and task performance of XAI information, and hence a determination on the output of the XAI design framework:

- The game will contain several visual tasks, depending on the number of independent variables chosen, with medium to high temporal demand.
- Participants are provided with AI decision support advice at a defined level of reliability.
- To fulfil the task, participants must decide whether to accept advice given by the AI or conduct the task manually.
- The game will require less than five minutes training
- The game will be suitable for participants with no previous experience in AI or decision support.
- The game must be able to be played on the Tobii T120 eye-tracker to collect the required gaze data.

It was also important that the complexity and time allowed are designed to maximize the impact of the AI advice and produce a benefit to trust in the AI while also allowing the task to be completed manually. To meet these requirements, an assessment platform designed around image classification 'Captcha' tasks, and error identification 'spot-the-difference' tasks was developed. These two tasks are widely understood, and representative of tasks associated with internet security and manufacturing applications. A prototype game containing both 'Captcha' and 'Spot the difference' tasks was developed and a pilot study on a small sample of 8 participants (5 male, 3 female) was conducted to ensure the requirements were met. The 'Captcha' image classification task, consisted of a 4x4 grid with an image in each grid square, some of which will contain targets of interest, some of which will contain objects similar to the targets of interest (decoys) and some will contain neither targets, nor decoys. Participants must decide how many squares contain targets within a limited time, and could consult an AI, both with and without XAI information. The ‘Spot the difference task provided two images of barcodes with differences between the two.

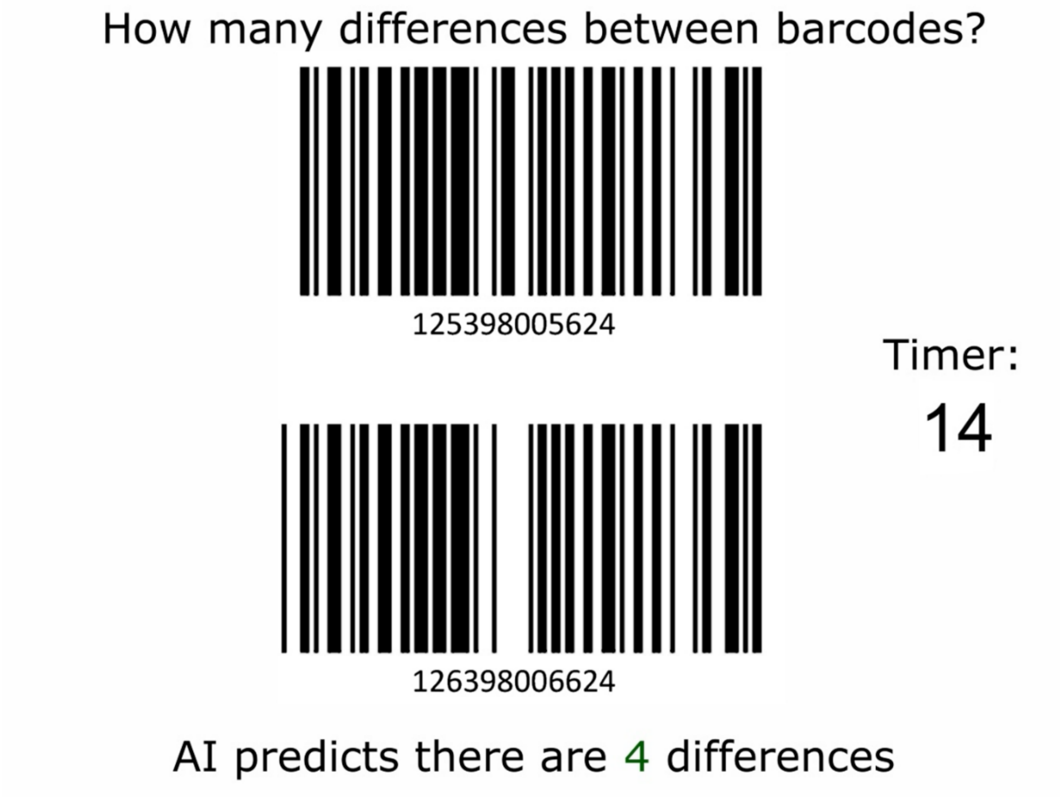
Detection and localization of differences between barcodes was selected as this is distinct from the image types in the Captcha task but still familiar to most potential users. Both tasks provided 20 seconds to complete before the images were blurred out, and an additional 10 seconds to enter their answer. The prototype hybrid tactical trust decision ‘Captcha’ and ‘Spot the difference’ are shown in Figures 3A.1 and 3A.2.



How many squares contain cats?

AI predicts 7 squares contain cats

Figure 3A.1 Baseline ‘Captcha’ task design


How many differences between barcodes?

AI predicts there are 4 differences

Figure 3A.2 Baseline ‘Spot the difference’ task design

The prototype game consisted of 16 tasks (8 of each type) with AI advice provided, along with a timer to enable tracking of game progress. To ensure consistency and definable system reliability, the AI advice was simulated. Baseline High and low reliability levels of 50 and 95% were adopted based on trust in automation studies conducted by Lu and Sarter (2019). To ensure there are multiple chances for incorrect AI advice, the high reliability task had 2 tasks with incorrect AI advice, where 8 out of 16 tasks contained incorrect AI advice in the low reliability AI tasks. This resulted in actual reliability levels of 50 and 87.5%. While these are different from those used in the Lu and Sarter study, they provide suitable separation from the 70% reliability threshold for human trust in automation defined by Wickens and Dixon (2007). This upper reliability level is in also an achievable level, in line with state of the art image classifiers such as

Florence (Yuan et al., 2021) or Meta Pseudo Model (Pham et al., 2021) which achieve ImageNet (Deng et al., 2009) classification accuracies of 90.5% (top 1), 99.02%(top 5) and 90% (top 1) 98.7% (top 5) respectively. To achieve high and low workload tasks, the number of targets, or differences were varied. In the 'Captcha' tasks, low workload tasks contained 2 to 4 grid squares with targets, where high workload tasks contained 6 to 8 targets; 'Spot-the-difference' tasks had 1 to 3 differences for low workload and 4 to 6 differences for high workload.

It was important that there was no ambiguity in correct answers, for instance, avoiding images that cannot reasonably be determined to contain a target, or where the target is too small to be reasonably identified given the screen resolution and viewing distance in the experimental setup. In the case of the 'spot the difference' task, differences should be unique differences in numbers, single additional elements, or single removed elements. Transposition of elements, or addition or subtraction of multiple adjacent elements, might be ambiguous to participants and result in impossible decisions or falsely trusting and mistrusting the AI.

In each of the 16 tasks, no two sets of images or barcodes were repeated to prevent learning effects. NASA TLX (Hart et al., 1988) and the adapted trust in automation scale (Jian, Bisantz, & Drury, 2000) were used to baseline participant workload level and intention to trust the AI, respectively. No XAI information was presented to participants in this phase.

The pilot study was conducted on the Tobii T120 Eye-tracker, following standard operating procedures (To and participants with limited or no prior experience of AI decision support systems, were able to complete the tasks at the viewing distance required for eye-tracking and with under 5 minutes of training. During the pilot study, participants could use the AI advice or conduct the tasks manually.

The NASA TLX scores gave mean temporal demand of the medium/low workload group as 9.7, SD 1.1 and 13.8 SD 1.0 in the high workload group, as measured by the NASA TLX. This difference in temporal demand was statistically significant, p=0.03. Similarly, the 50% and 87.5% reliability levels resulted in statistically significant intentional trust levels, using the adapted trust in automation scale (Jian, Bisantz, & Drury, 2000), with a low reliability mean trust of 3.71, SD 0.21 and high reliability mean trust 4.65, SD 0.19, p=0.01.

The pilot study established that the workload and reliability levels used in the pilot experiment would be suitable for the human-subjects evaluation. One area where there the pilot study was not successful was in developing behavioral trust in participants. There was no statistically significant difference between the high and low reliability AI, with participants trusting the AI advice less than 50% of the time in both models. This result did not align with the significant increase in intentional trust in the high-reliability AI tasks. A requirement was identified to produce an imperative for participants to match their behavioral trust to their intentional trust level. An additional requirement for the main phase of the trial was defined:

- The game will establish a benefit for participants who trust the AI.

The aim of this requirement was to establish a risk-reward trade-off, providing an imperative to trust the AI, and make the decision context more meaningful. This requirement establishes another requirement on the experimental design:

- The experiment must ensure that all participants have an equal opportunity to benefit from trust in the AI.

This concept and the associated requirements formed the basis of the assessment platform used in the human-subjects experiment.

Method

The experiment was conducted at the human performance laboratory, Dayton Campus, Wright State University. Where statistical significance tests are used, α=0.05. Where parametric statistical methods are not suitable, alternative non-parametric tests will be used.

Research Design

The experiment was a mixed design with the AI reliability level varied between subjects and all other independent variables varied within subjects, to maximize the information collected from each participant and without confounding results across automation reliability levels.

There are four proposed independent variables – two XAI Model specific and two task specific. The two model-specific independent variables are XAI information (4 levels: 0 - No XAI information, 1-reliability information, 2-confidence information, 3-reliability, and confidence information), and transparency timing (2 levels: before task, during task) in line with the TEXAS model (Stone et al. (2022). The task specific independent variables and their levels, as defined in the pilot experiment, are time criticality (2 levels: 1-TLX score 5 to 10, 2-TLX score 10 to 15) and system reliability (2 levels: high-87.5%, low-50%). This design results in design blocks, for XAI level, XAI timing, model reliability and design workload, which can be utilized in the analysis if there is significant variance between the blocks for workload and trust assessment. This gives a 4x2x2x2 experimental design with a total of 32 conditions. An experimental matrix was developed to reflect the combination of these independent variables, see Table 3A.1.

The high and low reliability AI models were achieved by introducing inaccurate AI advice to ‘break’ user’s trust. Both False Positive and False Negative instances of the tasks were included in the inaccurate task sets. To ensure each participant was only presented with the same tasks once and only once, 8 different versions of the ‘Captcha’ and ‘Spot the difference’ tasks were required (4 high reliability and 4 low reliability variants). Eight ‘Captcha’ tasks were developed with high and low variants for ‘Hawk’, ‘Motorcycle’, ‘Cat’, and ‘Flag’; target classes. Similarly, eight ‘Spot the difference’ tasks were generated, with high and low variants of 4 different barcode images. In each case, different target images or different variations between images were used to remove the potential for learning effects.

These task variants and the experimental matrix were combined, and a counterbalanced experimental design was developed. This counterbalanced design was modified to ensure that each participant was presented with inaccurate AI advice at the same point in the design. This ensured the experimental design did not inadvertently favor any participant, which was critical to ensuring fairness to all participants. The modified counterbalanced experimental design matrices for the high and low reliability experimental variants are shown in Tables 3A.2 and 3A.3.

Table 3A.1 Experimental Matrix for the XAI assessment study

Task related independent variables		**TEXAS model independent variables**		**Tas**
AI Reliability (between subjects)	**Workload (within subjects)**	**XAI Information (within subjects)**	**XAI timing (within subjects)**	
Low (50%)	Medium (5<TLX<10)	Level 0 (No XAI)	n/a	1
			n/a	2
		Level 1 (Rationale)	Before Task	3
			During Task	4
		Level 2 (Confidence)	Before Task	5
			During Task	6
		Level 3 (L1 and L2)	Before Task	7
			During Task	8
	High (10<TLX<15)	Level 0 (No XAI)	n/a	9
			n/a	10
		Level 1 (Rationale)	Before Task	11
			During Task	12
		Level 2 (Confidence)	Before Task	13
			During Task	14
		Level 3 (L1 and L2)	Before Task	15
			During Task	16
High (87.5%)	Medium (5<TLX<10)	Level 0 (No XAI)	n/a	17
			n/a	18
		Level 1 (Rationale)	Before Task	19
			During Task	20
		Level 2 (Confidence)	Before Task	21
			During Task	22
		Level 3 (L1 and L2)	Before Task	23
			During Task	24
	High (10<TLX<15)	Level 0 (No XAI)	n/a	25
			n/a	26
		Level 1 (Rationale)	Before Task	27
			During Task	28
		Level 2 (Confidence)	Before Task	29
			During Task	30
		Level 3 (L1 and L2)	Before Task	31
			During Task	32

Table 3A.2 Low reliability (50%) Counterbalanced experimental Design

Group	**Task ID:** **Task Accuracy: A (Accurate), FP (False Positive), FN (False Negative)** **Task type: Hawk(H), Motorcycle (M), Flag (F) Cat (C), Barcodes 1-4 (B1), (B2) (B3) B4)**															
1	1 H A	9 B1 A	3 B2 A	11 C A	5 M A	13 B3 A	7 B4 A	15 F A	10 B1 FN	4 C A	12 B2 A	6 B3 A	14 M A	8 F A	16 B4 A	2 H FP
2	3 H A	11 B1 A	5 B2 A	13 C A	7 M A	15 B3 A	1 B4 A	9 F A	12 B1 FN	6 C A	14 B2 A	8 B3 A	16 M A	2 F A	10 B4 A	4 H FP
3	5 H A	13 B1 A	7 B2 A	15 C A	1 M A	9 B3 A	3 B4 A	11 F A	14 B1 FN	8 C A	16 B2 A	2 B3 A	10 M A	4 F A	12 B4 A	6 H FP
4	7 H A	15 B1 A	1 B2 A	9 C A	3 M A	11 B3 A	5 B4 A	13 F A	16 B1 FN	2 C A	10 B2 A	4 B3 A	12 M A	6 F A	14 B4 A	8 H FP
5	2 B1 A	12 C A	4 B2 A	14 B3 A	6 M A	16 F A	8 B4 A	10 H A	9 H FN	1 B1 A	11 B2 A	3 C A	13 M A	5 B3 A	15 B4 A	7 F FP
6	4 B1 A	14 C A	6 B2 A	16 B3 A	8 M A	10 F A	2 B4 A	12 H A	11 H FN	3 B1 A	13 B2 A	5 C A	15 M A	7 B3 A	9 B4 A	1 F FP
7	6 B1 A	16 C A	8 B2 A	10 B3 A	2 M A	12 F A	4 B4 A	14 H A	13 H FN	5 B1 A	15 B2 A	7 C A	9 M A	1 B3 A	11 B4 A	3 F FP
8	8 B1 A	10 C A	2 B2 A	12 B3 A	4 M A	14 F A	6 B4 A	16 H A	15 H FN	7 B1 A	9 B2 A	1 C A	11 M A	3 B3 A	13 B4 A	5 F FP

Table 3A.3 High reliability (87.5%) Counterbalanced experimental Design

Group	**Task ID:** **Task Accuracy: A (Accurate), FP (False Positive), FN (False Negative)** **Task type: Hawk(H), Motorcycle (M), Flag (F) Cat (C), Barcodes 1-4 (B1), (B2) (B3) B4)**															
1	17 H A	25 B1 A	19 B2 A	27 C A	21 M A	29 B3 A	23 B4 A	31 F A	26 B1 FN	20 C A	28 B2 A	22 B3 A	30 M A	24 F A	32 B4 A	18 H FP
2	19 H A	27 B1 A	21 B2 A	29 C A	23 M A	31 B3 A	17 B4 A	25 F A	28 B1 FN	22 C A	30 B2 A	24 B3 A	32 M A	18 F A	26 B4 A	20 H FP
3	21 H A	29 B1 A	23 B2 A	31 C A	17 M A	25 B3 A	19 B4 A	27 F A	30 B1 FN	24 C A	32 B2 A	18 B3 A	26 M A	20 F A	28 B4 A	22 H FP
4	23 H A	31 B1 A	17 B2 A	25 C A	19 M A	27 B3 A	21 B4 A	29 F A	32 B1 FN	18 C A	26 B2 A	20 B3 A	28 M A	22 F A	30 B4 A	24 H FP
5	18 B1 A	28 C A	20 B2 A	30 B3 A	22 M A	32 F A	24 B4 A	26 H A	25 H FN	17 B1 A	27 B2 A	19 C A	29 M A	21 B3 A	31 B4 A	23 F FP
6	20 B1 A	30 C A	22 B2 A	32 B3 A	24 M A	26 F A	18 B4 A	28 H A	27 H FN	19 B1 A	29 B2 A	21 C A	31 M A	23 B3 A	25 B4 A	17 F FP
7	22 B1 A	32 C A	24 B2 A	26 B3 A	18 M A	26 F A	20 B4 A	30 H A	29 H FN	21 B1 A	31 B2 A	23 C A	25 M A	17 B3 A	27 B4 A	19 F FP
8	24 B1 A	26 C A	18 B2 A	28 B3 A	20 M A	30 F A	22 B4 A	32 H A	31 H FN	23 B1 A	25 B2 A	17 C A	27 M A	19 B3 A	29 B4 A	21 F FP

The dependent variables are trust, workload, and task performance. Trust will be measured using gaze duration and mean pupil diameter, baselined by the adapted trust in automation scale. The measure for gaze duration is the amount of time participants are looking at either the AI advice or XAI features during a task. Gaze entropy and mean pupil diameter will be used as the dependent variables for workload, baselined with the NASA TLX scale. Gaze entropy is calculated using the information entropy equation given in Equation 3A.1.

$$Entropy = H(X) = -\sum_{i=1}^{n} p(x_i) * \log_b p(x_i)$$

Equation 3A.1 Information entropy equation.

The number (n) and size in milliseconds) of the gaze duration bins for Equation 1 are derived from experimental data to ensure all gaze durations are captured in equal sized bins. The probability of a gaze duration length p(x_i) within a gaze duration bin is the number of gaze events that fall within that bin, divided by the total number of gaze events in a task. To scale gaze entropy between 0 and 1, b is set to the number of gaze duration bins. Task performance will be measured using the accuracy of the participant response and total task time.

Participants

Thirty-three participants (18M, 15F) were recruited from Wright State University. All participants were aged between 18 and 59 years old and able to read Arial, size 16 font on the Tobii eye-tracker screen, either unaided or with the use of corrective eyewear. No other exclusions were made in the recruitment process. To mitigate variations in the participants' inherent trust level, the experiment included baselining the participant's propensity to trust using the propensity to trust scale (Appendix C).

Equipment

The experiment was conducted on the Tobii eye tracker in the human performance lab. This is a binocular desktop eye tracker with a resolution of 1280x1024 screen, presented on a 16:13, 17" TFT display. It has a stated accuracy of 0.5 degrees, drift of less than 0.3 degrees, and a data rate of 60 Hz 120 Hz. Participants must maintain a distance of approximately 2 ft from the screen with freedom of head movement 30x22x30 cm.

Procedure

Participants completed the informed consent documentation and propensity to trust baselining (See Appendix C) on arrival in the human performance laboratory. Participants were then briefed on the use of an eye-tracker and in each case a calibration was performed. The participants then completed two familiarization tasks to ensure they were sufficiently familiar with the experiment and the tasks they were to perform. Each participant then completed the assessed tasks in line with experimental design. Participants completed the Adapted Trust in Automation Scale (Appendix B) and NASA TLX (Appendix A), after tasks 2, 6, 9, 11, 15 and 16. This provided baseline subjective assessment of user trust and workload for each XAI level and each of the inaccurate AI tasks.

Limitations

As this experiment was conducted using the university population, there was an age and knowledge bias in the population, where all but 2 of the participants were aged between 18 and 30. Valid eye detections for at least 50% of gaze durations, as classified by the Tobii eye tracker, were required for task data to be used in the eye tracker assessment. If a task had no valid gaze events, the entire task will be invalid from an eye-tracker analysis perspective and will be excluded from the analysis.

Experimental Assessment platform.

The information provided to the user is from a simulated AI, meaning the results are generated manually to ensure consistency. To provide realism, the simulated AI is based on a transfer learning implementation (Rosebrock, 2019) of an existing DL, CNN, VGG-16 (Simonyan and Zisserman, 2013). VGG-16 is an established and popular image classification model trained on 14 million images in the ImageNET dataset (Deng et al., 2009), with the ability to classify twenty thousand image types. The transfer learning approach allows the model to be tuned to perform better on specific user defined image sets through retraining of the model. As only the last layers of the model are retrained, the new model retains much of the existing image classification performance (Simonyan and Zisserman, 2013). The transfer learning approach has shown improvement in classification accuracy with a retrained, transfer learning model achieving 97% achieved on sample classifications of 90 Magnetic Resonance Imaging (MRI) images compared to baseline top 5 accuracy of the VGG-16 model of 92.9% (Keskar et al., 2016). This VGG-16 model is a Deep Learning image classifier, trained on 14 million images, with the ability to classify 20 000 image types.

The task design for the experimental platform for the analysis of the proposed XAI assessment measures was developed based on the initial concept hybrid tactical-trust decision game developed in the pilot study with the XAI versions of developed using the design framework (Stone et al., 2022)**.** Explainable AI information versions of the hybrid tactical-trust gamer were developed for each of the TEXAS model independent variables as defined in the experimental matrix, requiring six new XAI variants. The initial concept was used as the no XAI variant. For each of these XAI task versions, several prototype XAI schemes were developed, and down selected prior to the human-subjects testing. For conciseness, the detail of these concepts, and the XAI design framework process is not included in this paper.

Where XAI information provided before the task, it is provided before the AI has been run or provided any advice to the users. The primary function is to provide the user with an understanding of the type of model used, the underlying data and the potential for bias or error within one or either of these elements. XAI information provided during the task gives feedback on how the AI has made a specific decision and the probability that this is correct, based on the statistical model within the AI. Each of the concepts is explained in more detail in the following sections.

XAI Level 1 – Rationale (Before Task)

The chosen concept was based on providing an overview of the basis for AI decision-making while minimizing Jargon. The aim of this rationale is to give information on the complexity of the model but also to clarify that it is task optimized to look specifically for the targets of interest.

> "This AI image classifier is a pattern recognition model, trained on 14 million images, with the ability to classify 20 000 image types. The classifier was retrained and optimized to search for the images of interest in this task."

XAI Level 2 - Certainty (Before Task)

This concept is a presentation of the accuracy of the simulated retrained image classifier. Again, there is a focus on using interpretable language, without the need for specialized machine-learning knowledge.

> "This classifier has an overall test accuracy of 98% for classes in the target set. The false positive rate is 5%, while the false negative rate is 1%. This results in an overall mean advice accuracy of 88% when combined over 16 images"

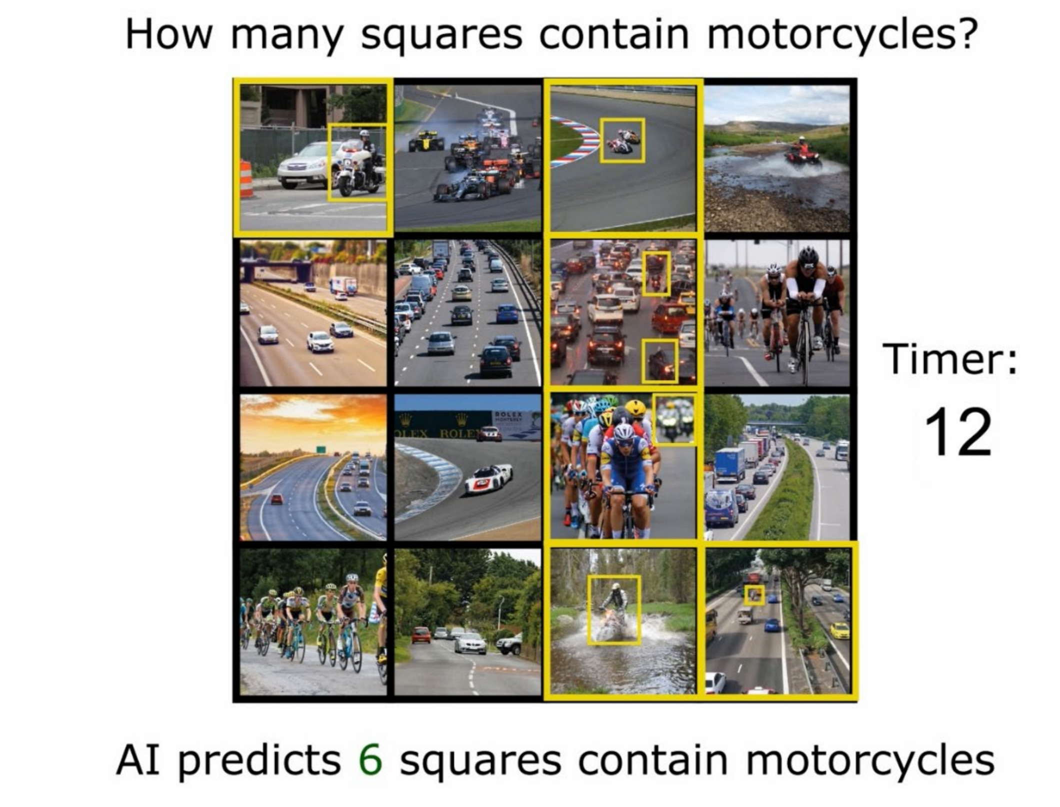
XAI Level 3 - Rationale and Certainty (Before Task)

Here both the individual rationale and confidence information are combined. The accuracy of the underlying model is compared human classification accuracy on similar datasets for additional context to the decision-maker.

> “This AI image classifier is a pattern recognition model, trained on 14 million images, with the ability to classify 20 000 image types. This is regarded as a state-of-the-art image classifier with an accuracy of 97%, compared with a human classification accuracy of 98% on the same image set. The classifier was retrained and optimized to search for the images of interest in this task. This classifier has an overall test accuracy of 98% for classes in the target set. The false positive rate is 5%, while the false negative rate is 1%. This results in an overall mean advice accuracy of 88% when combined over all 16 images”

XAI Level 1 - Rationale (During Task)

This concept provides the user with a classification status, by means of a yellow highlight for each box in the grid that the model predicts a target of interest is contained. Additional rationale information using a localization approach to highlight the specific targets of interest, the AI has identified within the image. Target localization is an increasingly popular solution in XAI, particularly for image classification where the model activations are used to highlight the locations within an image that are directly associated with the result, or ‘decision’ of the AI (Selvaraju, et al., 2017). Localization has been used to provide XAI information in a range of tasks from medical diagnosis (Zhao et al., 2021), autonomous vehicles (Grigorescu, et al., 2020; Johari & Swami, 2020) and airport security scanners (Zhang, 2019), and is well suited to both the ‘Captcha’ and ‘Spot the difference’ tasks in the hybrid tactical trust game. Localization provides a ‘quick-look’ for confirmation of true classifications by with highlighting the specific targets detected, but potentially could lead to over-trust in the system and a failure to interrogate the image for false negatives (López-Tapia et al., 2018). Other solutions such as target segmentation (Tseng et al., 2021; Vitulano, et al.,1995). and activation heatmaps Selvaraju et al., 2017) were considered, but localization provides less obscuration of the target than the heatmap solution and does not obscure the non-target areas as image segmentation algorithms might. An example of the level 1, rationale XAI information task is shown in Figure 3A.3. All images used in the generation of these tasks were stock images with no attribution required.



How many squares contain motorcycles?

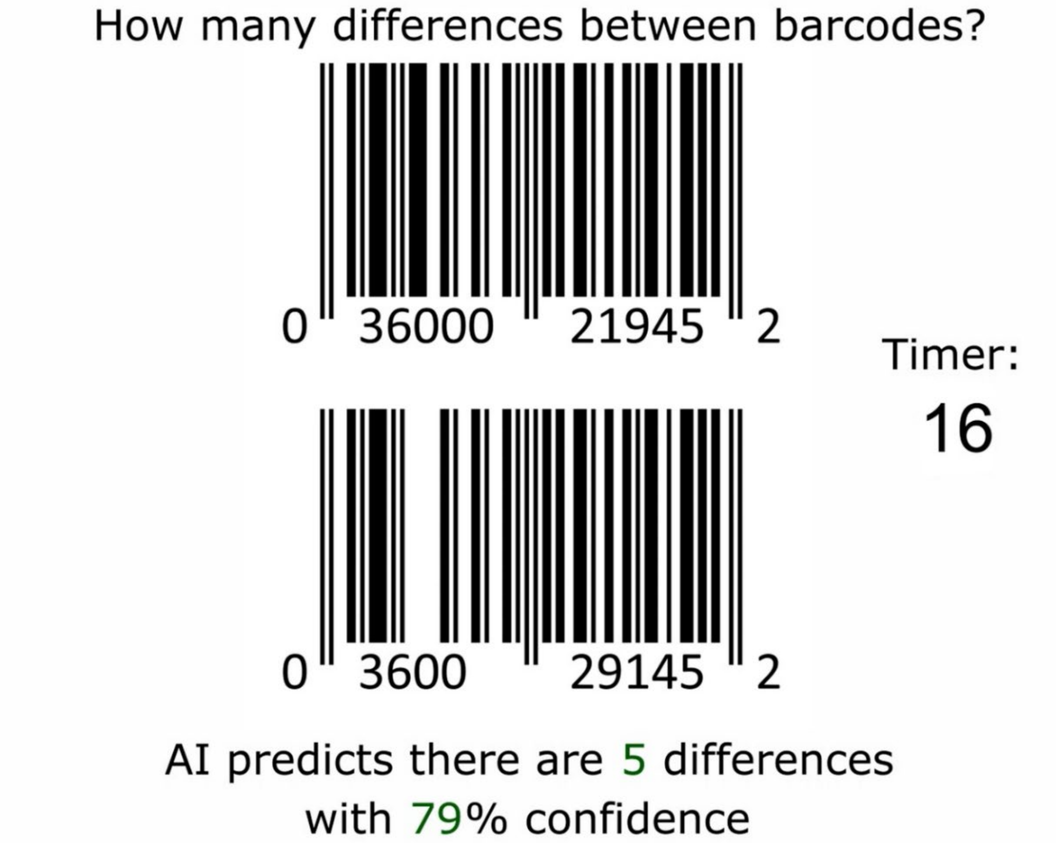
Timer:
12

AI predicts 6 squares contain motorcycles

Figure 3A.3 Example of the Rationale XAI information task

XAI Level 2 – Certainty (During task)

A concept with a simple confidence description, based on a simulated result of an application of the AI model was chosen. This provides a quick-look confidence in the AI advice and allows the user time to conduct a manual review of the task grid if the probability is deemed low by the user. The word ‘confidence’ is used, rather than probability or certainty, as this was the intended interpretation of the intention of the AI advice, see Figure 3A.4



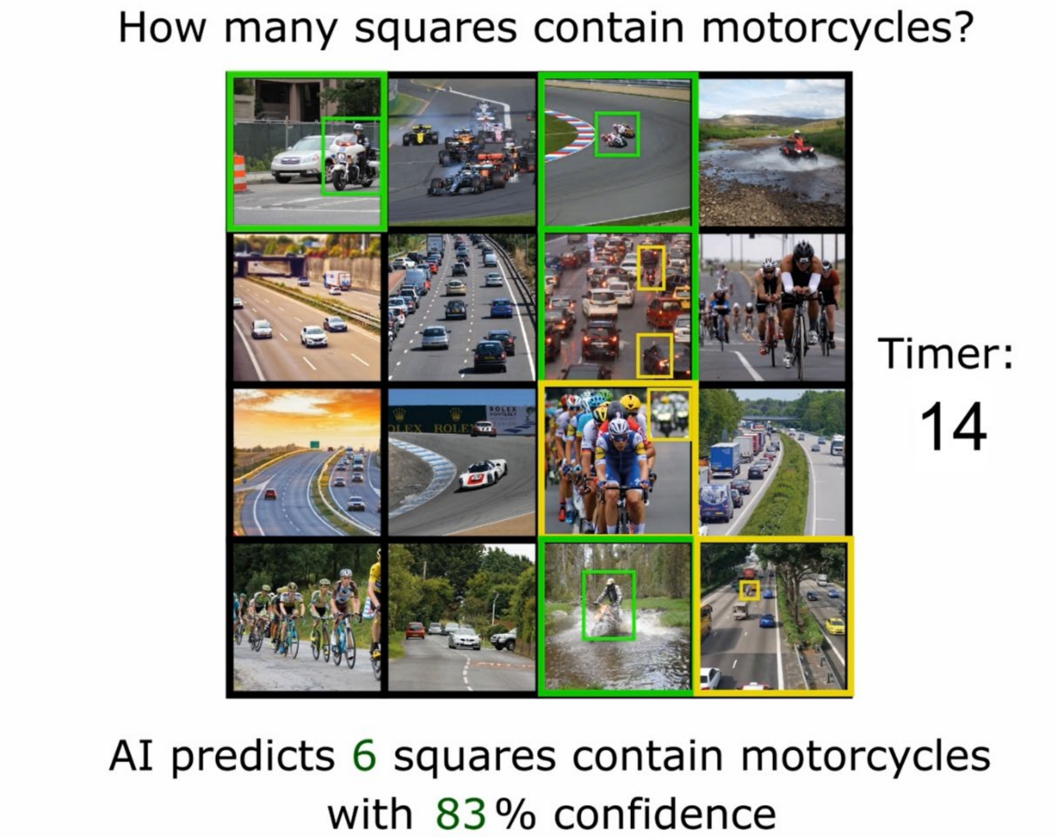
How many differences between barcodes?

AI predicts there are 5 differences with 79% confidence

Figure 3A.4 Example of Certainty XAI information task

XAI Level 3 - Rationale and Certainty (During Task)

This concept combines both the rationale and the confidence task XAI information but introduces color-coding of the classification and localization lines. High confidence classifications are shown in green and medium confidence in yellow. Indications in red are low-probability and not classified as targets by the AI but considered borderline in probability terms. However, the inclusion aims to provide additional information in the detection of potential false negatives. The exact thresholds in an implementation will depend on the model accuracy but are not presented to the participants in this study. The box classification highlights, which refer to the categorization of the box as containing a target are also color-coded. In this case, the color is defined by either the highest probability of a target classification within a box, or where the cumulative classification confidence reaches the predefined high or medium threshold, see Figure 3A.5.


How many squares contain motorcycles?

Timer:
14

AI predicts 6 squares contain motorcycles
with 83% confidence

Figure 3A.5 Example of Rationale and certainty XAI information task

Each of these designs was selected based on a design review and heuristic evaluation, however the intention is not that these are necessarily the best solutions, but more to achieve a breadth of designs to test the six new XAI information task sets.

The output of an implementation of this test method could be to distinguish between any two of the initial concepts using the evidence-based assessment method developed in this paper.

Intention to Trust Requirement

To develop the trust imperative, there is a need to include a risk and reward in what participants choose to do with the information from the AI. Including rewards for correct action and penalties for following incorrect advice or failing to follow correct advice, introduces a trust paradigm similar to the prisoner dilemma game (Poundstone, 1993).

Ferrer and Farolfi (2019) note that Trust is revealed when an agent performs an initial sacrifice, that is, an action which, depending on the reaction of another agent, might be detrimental to the first agent's own interests. You put yourself in somebody else's hands, trust is repaid, and the second agent is revealed to be trustworthy, if his or her reaction offsets and compensates for the first agent's sacrifice. This is an important feature to include in the trust-based tactical decision. As such, there is a need for a small reward or prize, which the user has a greater chance of winning if correct advice is followed and a greater chance of losing if either incorrect advice is followed or correct advice is not followed. In this experiment, a $25 prize was determined for the best performing participant. In this case, the performance was judged on the minimum time to complete the tasks. To provide a risk element, a 20s time penalty is introduced for each incorrect answer. As a result, there is a benefit to trusting the AI, but only when the AI is correct. In this experiment, the simulated AI provides advice within 2 seconds, considerably faster than would be possible for the participant to complete the task without AI advice. There is, therefore, a benefit to trusting the AI, but to be absolutely correct the participant would need to be able to quickly calibrate their trust in the AI and determine if they believe the advice or not, before choosing whether to manually conduct the task.

Research Questions

The following research questions were defined to be answered in order to meet the research objectives:

1. Are there reliable physiological predictors of workload?
   a. Is gaze entropy a predictor of workload?
   b. Is pupil diameter a predictor of workload?
2. Are there reliable physiological predictors of trust in automation?
   a. Is gaze duration a predictor of human trust in automation?
   b. Is pupil diameter a predictor of human trust in automation?
3. What are the relationships, and effect interactions, if any, between XAI, task performance and these measures of trust and workload.

There is an implicit question arising from research questions 1 and 2 that is of fundamental importance to this study:

4. Are the proposed measures for workload and trust confounded in time-critical tasks?

The research question arising from the secondary objective is:

5. Does the XAI design developed using the XAI design framework (Stone et al., 2022) improve performance and trust while reducing workload, and does it result in a usable design.

Hypotheses

The following hypotheses were derived from the research questions to provide the basis for the statistical analysis to be completed in this study. The alternative hypotheses in all cases, except hypothesis three are directional in line with predictions made by the original authors.

Hypothesis 1a:

- H10 Gaze Entropy is not affected by workload.
- H1a Gaze entropy increases with increased workload.

Hypothesis 1b:

- H10 Mean pupil diameter is not affected by workload.
- H1a Mean pupil diameter decreases with increased workload.

Hypothesis 2a:

- H10 Gaze duration is not affected by trust.
- H1a Gaze duration decreases with increased trust.

Hypothesis 2b:

- H10 Mean pupil diameter is not affected by trust.
- H1a Mean pupil diameter decreases with trust.

Hypothesis 3:

- H30 There are no interactions or relationship between the main effects.
- H3a1 Mean pupil diameter decreases with workload.

There are no formal hypotheses for research question 4, the answer will be derived from Hypotheses 1a, 1b, 2a and 2b. Research question 5 will be answered by means of a systematic usability study (SUS) (See Appendix D), Brooke (1986), delivered to each participant upon conclusion of the study, specifically on the usability of the XAI information. There is no statistical test associated with the SUS evaluation.

Analysis

The effect of independent variables will be assessed in line using null hypothesis significance tests, specifically, t-tests and Chi-squared tests as appropriate. For

Hypotheses 1 and 2, interaction effects will also be considered, as these are also potentially significant (Deng et al., 2009). If required, non-parametric testing will be conducted. Where parametric statistical methods are not suitable, alternative non-parametric tests will be used.

Required Participants

The minimum number of participants required for an experimental power of 0.9, using $\alpha=0.05$ was determined by the higher of the participant number calculations for eye-tracker trust and workload calculations from the pilot study. In this case, $\mu_1=9.8$, $\mu_2=13.1$, and the common standard deviation was $\sigma=32.5$. This resulted in a minimum number of participants of 13. As there are two separate experimental groups, this resulted in a requirement for 26 participants. To have 2 complete repetitions of the counter-balanced experimental design, a minimum of 32 participants were required.

Results

Data Validity and exclusions

33 participants each conducted 16 visual tasks, giving 528 tasks completed. The high and low reliability tasks were conducted between subjects, the remaining independent variables were tested within subjects with all participants exposed to each level. There was one additional participant in the high-reliability set, giving 272 tasks, versus 256 tasks in the low-reliability set.

- Of these tasks, 513 gave good validity eye-tracker results, as per the validity criteria defined in the procedure. Leaving 15 (2.8%) with poor quality data, classed as invalid and not to be used in the analysis of eye tracker data. These were split across the experimental conditions as follows: Of the 528 tasks, 15 were invalid. There were 10 (3.9%) invalid low reliability tasks, meaning 5 high reliability tests were invalid (1.8 %).

- The low workload tasks included 7 classified as invalid (2.7%), where the high reliability tasks had 8 invalids (3.0%)
- There were 3 invalid tasks with no explainable AI information (2.3%); the tasks with level 1 XAI information (XAI rationale only) had 6 invalid task (4.5%); level 2 tasks (XAI confidence only) had 5 invalid tasks (3.8%); and the level 3 XAI information tasks only had one invalid result (0.75%).

The tasks where accurate AI advice was given had 8 invalid tasks (2.2%) and false positive and false negative tasks had 4 (3.6%) and 3 (6.1%) invalid tasks respectively. No experimental condition had more than 6.1% of the associated tasks classed as invalid. This does not diminish the validity of the overall experiment with respect to statistical power and the required participant numbers as more than the minimum 26 were recruited. Where eye-tracker data is used in the following analysis, invalid tasks were excluded.

Eye Tracker Workload Assessment

To calculate gaze entropy, it was necessary to count the number of gaze durations in specific bins. The maximum gaze duration was 2000ms and gaze counts were assigned to 40 x 50ms bins, covering 0 to 2000ms. This ensured all fixations were assigned to equal bins.

XAII Gaze entropy significantly increased creased with increasing workload p=0.04. However, this simple model only had an R^2 of 0.05. The independent variables were used as blocking variables, to account for more variation in the data. A reverse optimization model was constructed, however none of the blocking variables were significant and no improvement to the R^2 was achieved. The relationship between mean pupil diameter was also considered but no significant effect or interaction effect between pupil diameter and gaze entropy was observed.

Eye Tracker Trust Assessment

The intention to trust as measured by the adapted trust in automation scale (Jian, Bisantz, & Drury, 2000), was normalized according to the participants propensity to trust baselining for this analysis. Gaze Duration was significantly longer on low AI reliability tasks $p < 0.001$, $R^2 = 0.11$. Gaze duration significantly decreased with increasing trust $p=0.03$, however, this simple model only had an R^2 of 0.01. The independent variables were used as blocking variables, to account for more variation in the data. A reverse optimization model was constructed and the significant blocking effect that remained was design accuracy (accurate, FP, FN) $p=0.011$. The resulting model increased the significance of the relationship between gaze duration and intention to trust to $p=0.01$ and an R^2 of 0.1.

The relationship between mean pupil diameter and trust was also considered and found to decrease significantly with increasing trust $p=0.03$, but again had a low R^2 of 0.02, A model combining the XAI gaze duration, mean pupil diameter, and design accuracy was built, resulting in a model where all three effects were significant - Gaze duration ($p=0.001$), design accuracy ($p=0.01$) and mean pupil diameter ($p=0.02$). This model had an increased R^2 of 0.12 and R^2 adjusted of 0.1.

Behavioral Trust

Behavioral trust is where the participants' actions are to implement the given by automation. The trust element is implied by the action, as we cannot be sure there was trust or if the participants came to the same conclusion without trusting the AI. Participants exhibited behavioral trust in the AI occurred in 286 of the 528 valid tasks, or at a rate 50.9%. The high-reliability tasks showed higher behavioral trust with 156 trusted tasks out of 272, giving a trust rate of 57.3%, whereas, in the low-reliability tasks, only 130 of the 256 tasks, or 50.8% indicate behavioral trust. When XAI was considered,

Relationship between trust, workload, and task performance

Logistic regression analysis was conducted to determine if there were significant relationships between the task outcome and both trust and workload. In both cases, chi-square tests showed significance in the relationships, with increased workload reducing task performance, p<0.001, R2 =0.05 and increased trust improving task performance, p = 0.0175, R2 =0.02.

- There was a significant negative relationship between improved task accuracy and increased workload, p<0.001, R2 =0.05.
- The logistic model of workload v task accuracy had an accuracy of 60%.
- There was a significant positive relationship between improved task accuracy and increased trust, p = 0.0175, R2 =0.02.
- The logistic model of the trust v task accuracy had an accuracy of 58%.

XAI design performance

Across all tasks, there was a slight increase in correct outcomes associated with the level 3 XAI information, but this was not statistically significant, see Figure 3A.6.

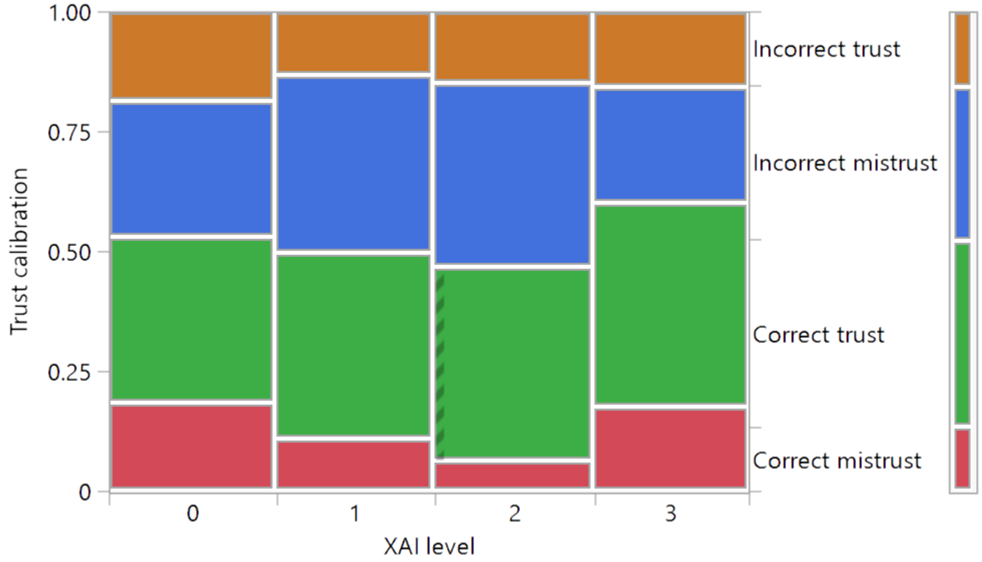


Figure 3A.6 Trust Calibration outcome by XAI level

When the results were broken down by the design accuracy variable, there was a similar improvement in correct outcomes in the accurate AI information tasks, see Figure 3A.7.

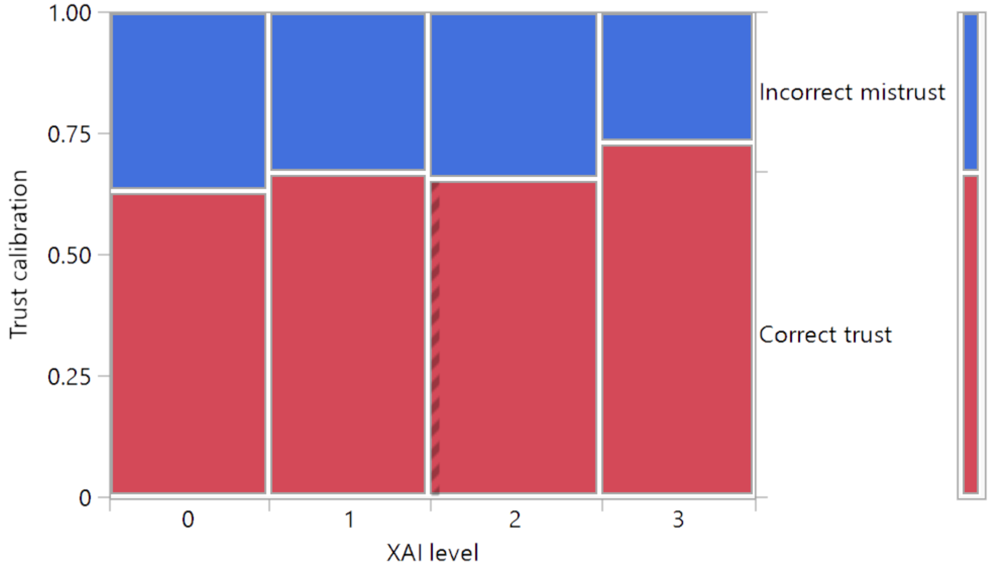


Figure 3A.7 Trust Calibration outcome by XAI level (Accurate AI)

However, in the false negative tasks, the XAI tasks showed worse outcomes than the no XAI tasks, see Figure 3A.8.

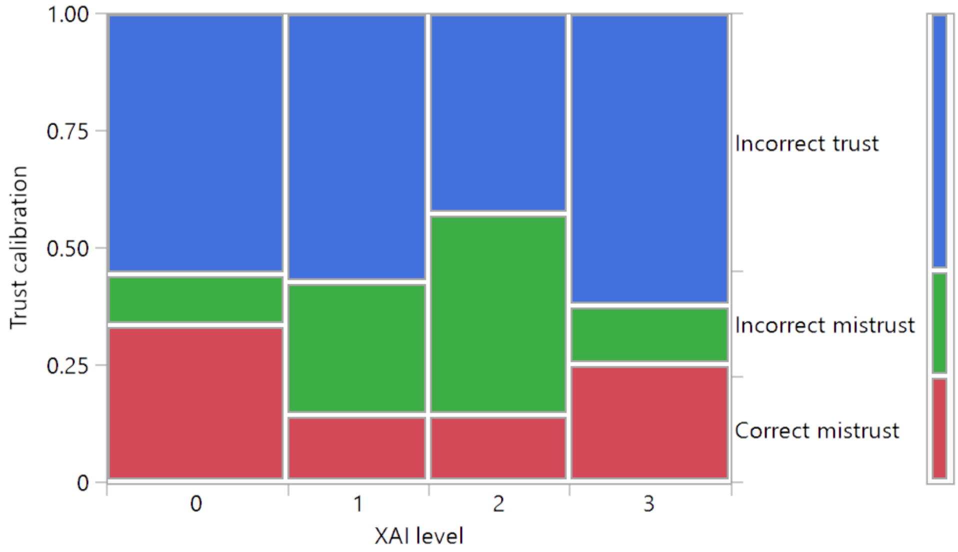


Figure 3A.8 Trust Calibration outcome by XAI level (False Negative AI)

When the results of the false positive tasks are examined, this trend is reversed, and the Level 3 XAI system showed slightly improved performance, see Figure 3A.9.

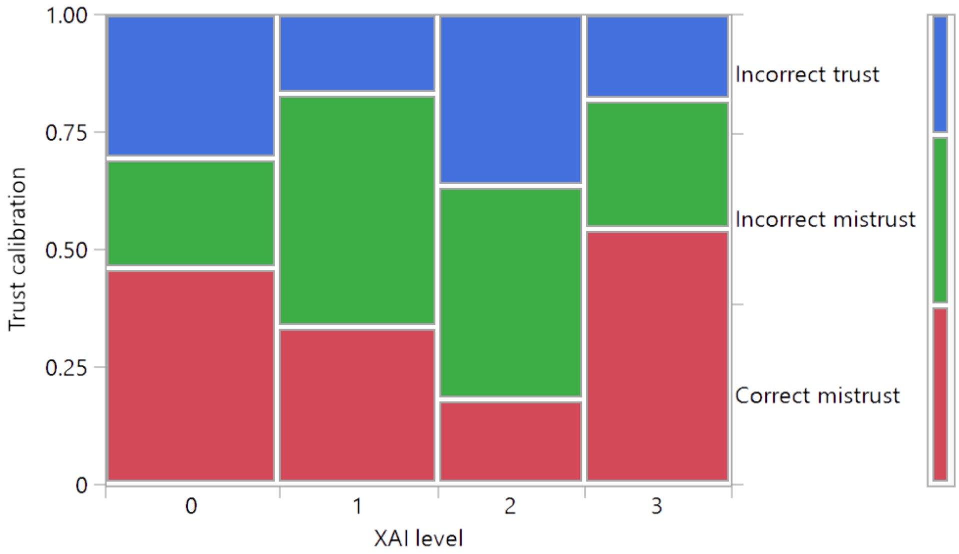


Figure 3A.9 Trust Calibration outcome by XAI level (False Positive AI)

None of these effects were statistically significant, however design accuracy had a significant effect on the task performance at all four XAI levels. The result for the level 3

XAI highlights that this XAI design is performs much better on false positive, than false negative tasks, See Figure 3A.10.

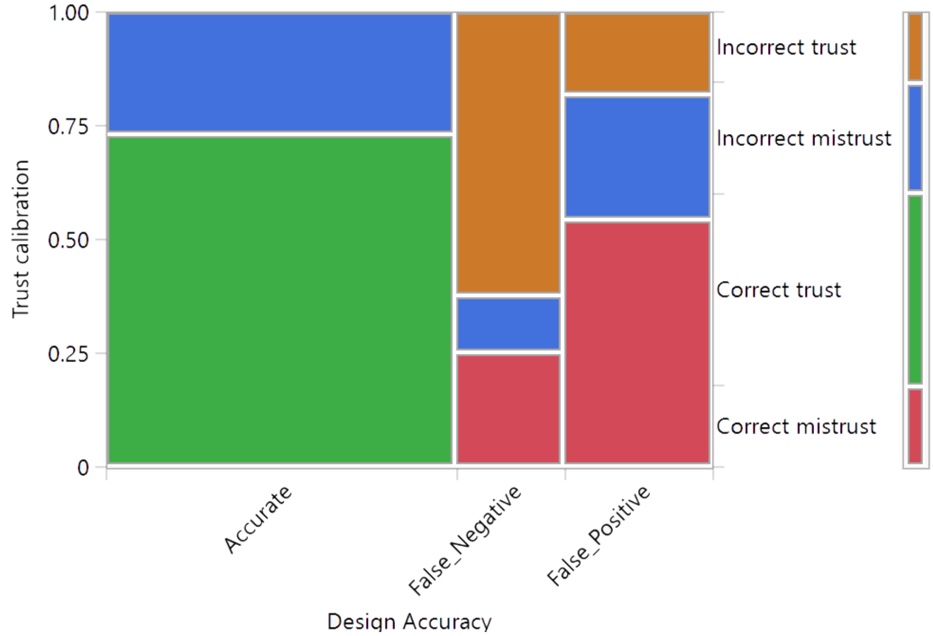


Figure 3A.10 Trust Calibration outcome by designed AI accuracy (XAI level 3)

SUS Score

The mean SUS score for the XAI was 53.8, which rates below the 50th percentile score of 68 and is classed as marginally usable according to Lewis (2018), or equivalent to the 15-34th percentile.

Discussion and Conclusions

The following conclusions relate directly to the research questions. The following conclusions relate to research questions 1 and 2:

- The eye-tracker results showed gaze entropy to have a significant relationship with workload, p=0.04. however, the low R2 of 0.05, means that this does not account for much of the variation in the model. There is some evidence to support gaze entropy as a predictor of workload in TCS, but further studies are required.

- There was no statistically significant relationship between pupil diameter and workload.
- The eye tracker results showed gaze duration to have a significant correlation to trust, p=0.03, R2 =0.01 and this effect was more pronounced when accounting for the design accuracy of the task, p=0.001, R2 =0.1. There also is evidence to support gaze duration as a valid predictor of trust in TCS.
- The eye tracker results showed pupil diameter to have a statistically significant relationship with intention to trust, p=0.03, R2 =0.1, There is evidence to support pupil diameter as a predictor of trust in TCS.
- When pupil diameter and gaze duration were combined in a predictive model, they both had a statistically significant relationship with intention to trust, p=0.02 and p=0.001, respectively, and this model had an increased R2 of 0.12 and R2 adjusted of 0.1. There is evidence that this combined model is a valid predictor of trust in TCS.
- Although both gaze entropy and gaze duration were found to be statistically significant predictors of workload and trust respectively, statistical analyses show very low R2 values, indicating this only accounts for 10% of the variation. While some of this variability can be explained by the number of variables in the test and innate human variability, there is a need to understand this further and develop models that account for more of the variation in the result to ensure that these metrics can be used with confidence.

The following conclusions relate to research question 3, regarding the interactions between trust, workload and task performance

- It is concluded that in TCS, improved trust can improve task performance, while increased workload can negatively impact task performance.
- The XAI information did not result in statistically significant differences in task performance between each other, or the no XAI tasks.

The following conclusions relate to research question 4, regarding the ability to concurrently measure trust and workload using the eye-tracker.

- Both trust and workload could be measured concurrently in the time-critical system, indicating that these measurements do not confound each other, even in the case of high temporal demand.

The following conclusions relate to research question 5, regarding the performance of the XAI instantiation delivered by the design framework.

- There was no statistically significant difference in behavioral trust between the XAI and no XAI tasks. This indicates that the XAI did not encourage trust calibration as intended. More generally, the behavioral trust of the participants matched the low reliability AI tasks well but did not increase to match the reliability of the high reliability tasks.

The XAI framework delivered an instantiation, which was usable on first iteration but did not provide improvements in task performance, workload or trust. Where the XAI framework showed promise was in the ability to quickly assess designs and feed information back into the iterative design process. Looking at the results more closely, analysis of the results by AI accuracy showed that the level 3 XAI significantly improved task performance in the case of false positive AI advice, but significantly reduced task performance in the case of false negative advice. This indicates a tendency to over-trust the XAI information, and that the system is exacerbating over-reliance in the systems. The SUS score for the XAI usability suggests that participants found the design usable had some difficulty in interpreting the XAI. This might indicate the initial instantiation of the XAI framework has room for improvement but also indicates the inclusion of the SUS score in assessment of XAI designs can provide useful information to designers, where other forms of usability assessment such as root cause analysis are not possible due to the time-critical nature of the task.

The aim of this study was to develop evidence for objective, physiological metrics for trust and workload for XAI in TCS. This was demonstrated in the results, although there is a great deal of variation in the data that remains unaccounted for. This variation was also noted in the baseline measurements and may be characteristic of the variability of human-subjects in this specific context. Human variability is complex and context dependent (Smith et al., 2014), so it is hard to conclusively determine this variability as down to the human participants without further exploration of the problem at hand. It is also important that these metrics align with the XAI measurement requirements and provide consistency of approach throughout the XAI design framework.

The variability of the results may also be a result of the experiment being underpowered. The calculations for the minimum number of participants were based mean values for trust and workload obtained in a pilot study. The sample size of this study was only 8 participants so the values used in the power analysis might not have been representative, causing the power analysis to be erroneous or for the real mean trust and workload values to have been closer than desired.

Although not an absolutely conclusive result, the findings were in line with expectations, and provide some of the evidence for the validity of eye-tracker measurements the assessment of trust and workload, and that these can be measured simultaneously.

The provision of objective, physiological metrics allows the researcher to include more tasks in an equivalent study period by removing the requirement for the participants to complete subjective assessment questionnaires. Objective results, while still subject to variability, are less dependent on the participant correctly interpreting the meaning of the test and potentially less subject to individual and systematic experimental bias. (Pronin, Lin, & Ross, 2002).

The confirmation of the expected mapping between trust, workload, and task performance provides a feedback loop for designers and enables better prediction of the potential impact of design changes on task performance. Again, there are potential issues with the lack of variability accounted in the models.

The results of the XAI information instantiation were less positive, with little or no impact on trust calibration and only marginal usability ratings. This does however demonstrate that the measures developed, and the hybrid tactical-trust game assessment platform, can be used in the assessment of XAI. In this case, alternative forms of XAI could be tried such as presenting a dynamic trust calibration level to the participant or introducing more feedback on correct and incorrect behavioral trust.

There is also evidence in this case that the XAI instantiation, while marginally improving overall performance, induces over-reliance in false negative tasks. While this may appear to be a negative result, it is confirmation that the assessment method defined in this study allows the designer to determine how well the XAI information performs in each of these cases and make adjustments in the next iteration of the design. In this case, the XAI information could be modified, or the AI itself could be adjusted to produce fewer false negative results. While this might produce more false positives, the predicted overall accuracy of the new model could be compared to the results across the 3 levels of design accuracy.

One finding of this study, that presses a need for further investigation, is the lack of behavioral trust in experimental participants in the AI. Instead they often choose to rely on manual completion of the tasks, despite being both less accurate and considerably slower than the AI. This effect was present in both XAI and AI results, and potentially influenced the results of the XAI instantiation assessment. This difficulty in getting explainability or transparency to shift behavioral trust was also noted by Schmidt et al., (2020), who also highlight the potential improve human performance is not always enough for humans to trust an AI.

The study included provision for additional incentive to trust the AI, from highlighting the benefits of trust to providing financial incentive to the participant in the form of a prize for the best performer this did not significantly impact the intentional trust in participants. This finding suggests that either the incentive provided was not enough to induce behavioral trust between the participants and the AI, or that the participants were not willing to trust the AI at all in this task. While the drivers of this are not clear, this result demonstrates the difficulty in achieving behavioral trust in users, particularly in short assessment studies, where there is limited time to build trust. This may be an impediment to XAI design studies and highlights that this is an element of the framework, and assessment platform development that requires careful consideration.

Please see the Reference section at the end of the document

End of Paper: Development of a Novel Hybrid Cognitive Model Validation Framework

for Implementation Under COVID-19 Restrictions

Chapter 4 – Contributions and future work

In this research, I establish a design framework for XAI in TCS. This framework was developed from an understanding of human-AI teaming, through the development and validation of a cognitive model for a complex human-machine task, along with a Deep learning model aligned to the same problem. The final two studies detailed in this research establish the design thinking framework for human-centric explainable artificial intelligence in time-critical systems, which is the key contribution of this research, see figure 4.1.

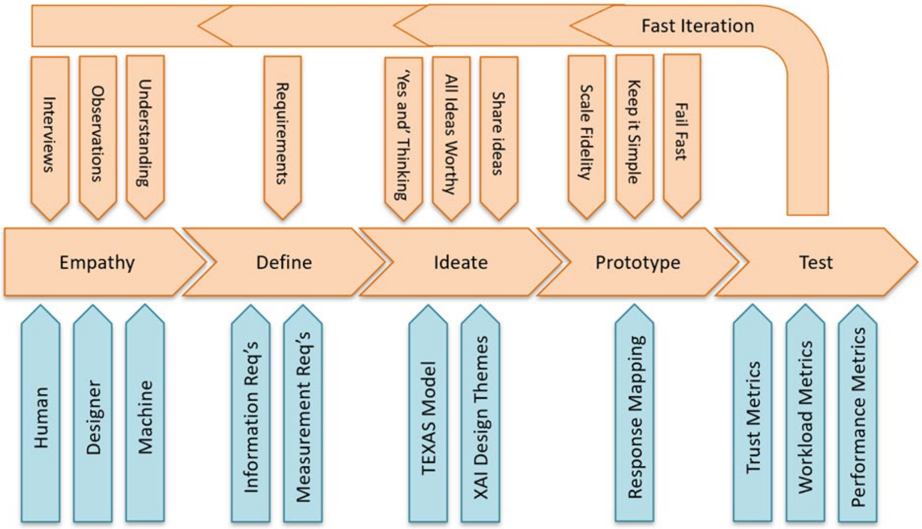


Figure 4.1. A Design Thinking Framework for Human-Centric Explainable Artificial Intelligence in Time-Critical Systems

This framework incorporates XAI specific elements at each stage but preserves empathy as the foundation of the design problem, along with the open sharing of ideas and fast iteration of solutions that characterize design thinking. In the define stage, I establish the importance of trust, workload, and task performance as drivers of the XAI design, along with specific information requirements for XAI. This framework allows AI developers or user experience and user interface designers, who may not be familiar with the concepts and drivers of XAI to quickly understand the problem and provides specific tools in the form of the TEXAS model, XAI response mapping and objective, physiological performance assessment measures for both trust and workload. These metrics are aligned to XAI measurement requirements and form a consistent focus through the XAI design framework. The use of physiological measures eliminates the reliance on subjective questionnaires and reduces the total task time in XAI studies. This in turn increases the total number of tasks that can be conducted in the assessment of XAI in TCS. The results of the human-subjects experiment show eye-tracking measures have validity as metrics for workload and trust, but given the low R2 values, there is a need for further research.

The use of eye-tracking to measure trust and workload simultaneously has broad application beyond XAI, in any system where human trust in automation, or workload are desirable assessment metrics, and particularly in time-critical tasks, where I have demonstrated the effectiveness of eye-tracking measures for concurrent assessment of trust and workload. The use of eye-trackers to measure trust and workload is less intrusive than questionnaire-based assessment and provides real-time measurement, which could benefit any research where trust and/or workload are desirable measurements.

The study provided evidence for the relationship between task performance, workload and trust in TCS. Demonstration of this relationship feeds directly into the iterative prototype-test phase of the XAI design framework, allowing designers to predict the impact of design decisions more confidently in the iterative phase of design thinking. More broadly, this can be used in any system where task performance is contingent on workload and trust to guide iterative prototype development. This broad directional relationship between workload and trust is dependent on reliability and temporal demand of the system, establishing a model to determine measures of individual differences in these responses would be a useful addition to the XAI design framework assessment toolbox and allow more confident prediction of the potential impact of design changes at the prototype stage.

The assessment platform developed in this research is a hybrid tactical-trust decision game. This game was developed as no assessment platform combining task-based tactical decision making with a trust requirement was found. This study establishes the core requirements of a hybrid tactical-trust decision game, for the assessment of XAI in visual decision tasks, and the use of 'Captcha' and 'Spot the difference' tasks as options for generic tasks for the assessment of XAI designs. The significance of this contribution, and the design framework for XAI in TCS specifically, is to ensure that XAI does not come at the cost of task performance and that unintended consequences can be better predicted and avoided.

While these are significant contributions, they are all still relatively new approaches and there are some important knowledge gaps that remain. Although the results show good correlation between baseline workload and trust metrics and the experimental physiological measures, there was still a lot of variation unaccounted for in the results. This means the output metrics are best used for comparative human-subjects analysis, as there is too much variation for individual results to be predictive of high or low workload or trust. Further investigation of the drivers of this variation is required to determine if it can be accounted for either in the available data from the eye-tracker or through greater experimental control. There is a possibility, as with any human-subjects research that this variation is related to innate human response variability and will exist in any model or metric for trust and workload.

The tactical-trust decision game showed promise, but further investigation is needed to establish a means to account for the lack of trust in the AI. Developing an assessment platform that can develop behavioral trust in participants is important in the assessment of XAI design. This was potentially one of the reasons it was difficult to determine any change in behavioral trust, or indeed in the performance of the XAI information in the final study. While the XAI design framework can function without an improvement in the behavioral trust of the assessment platform, providing this could result in faster, more accurate assessment of the impact of XAI.

Situational Awareness is an important construct to consider in the development and evaluation of XAI in TCS. In this research, SA was initially considered as a measure to be included in the design framework, however the preliminary investigations highlighted the difficulties with collecting subjective SA data in TCS and no alternative physiological measure was identified. As a result, SA was not considered in the final stage of this research, but remains an important consideration in the evaluation of XAI integration. Development of appropriate means to measure SA in TCS should be considered in future studies to provide a design framework with comprehensive performance assessment capabilities.

In addition to these design thinking elements, the concept of trust calibration was an important part of XAI, that was potentially under-represented in the design framework. The ability to present the user with their trust calibration state may also be a means to improve behavioral trust in the users, as it provides direct feedback on how much they are trusting the AI. Identification of potential metrics to assess trust calibration is a key requirement if the trust calibration status of users is to be provided as part of the XAI information. Further research into this is required, especially if the application of the XAI is for operational purposes, where a baseline for correct decisions, and hence behavioral trust is not available to the designer. In this case, the only option would be to provide the user with their trust rate, without any confirmation as to whether the AI, and the decision to trust it was correct.

The high variability in the results was potentially associated with the experiment being underpowered. The experiment could be repeated with either more participants or adjusted levels of the independent variables of temporal demand and AI reliability to increase experimental power and determine if this is a contributor to the low confidence in the results of the final experiment.

The final area that I want to discuss for consideration for future research is developing metrics for equity to determine the performance of the empathy phase. This is an element of research I considered during this research but did not take forward. The principal areas of development to focus on are the user and the internal Biases of the designer. AI for social good, and equitable AI (Gilbert, 2021) are important concepts that should be considered in the design of all AI systems. In establishing their vision of Equitable AI, Gilbert (2021) aims to develop AI to create holistic diversity. The need for equity in AI is part of AI for social good (Tomašev, Cornebise, and Hutter 2020), along with explainability of AI. While equity is not a requirement in the design of explainable systems, failing to consider it could be seen as a failure of the core aims of XAI. Providing specific elements regarding equity in the empathy stage can cue the designer to consider both users and those who might be affected by the systems.

Foulds, et al., (2020). Propose an intersectional definition of fairness, based on their differential fairness measure, derived from the 80% rule, established in the Code of Federal Regulation (CFR, 1978). More specifically, if the ratio of probabilities of a beneficial outcome, between a disadvantaged and an advantaged group, is less than 0.8, there is said to be legal evidence of an adverse impact (Foulds, et al., 2020). This is formalized in equation 4.1.

$$P(M(x)=1|group\ A)/P(M(x)=1|group\ B)<0.8.$$

Equation 4.1. Formalization of the 80% discrimination rule (Foulds, et al., 2020).

Where the deep understanding of the user and underlying data available for the AI indicate the possibility of reduction in beneficial effects that approach 80% of the baseline population, at a minimum, the XAI implementation should highlight this.

Furthermore, ensuring this is achieved across the full spectrum of diversity is equally important. Artificial Intelligence is trained using data generated by society, and therefore reflects the biases inherent in society (Bauer & Lizotte (2021), and Ciston (2019) highlights the need for intersectionality in AI from data to design through to the final implementation and considering all levels of user. Intersectional theory was first established by Crenshaw (1989) and identifies the relationships between issues of sexism and racism and establishes the need to consider all oppressive structures, and the relationship between the negative impacts of them all. The key demographics identified in intersectional theory are gender, race, ethnicity, sexual orientation, gender identity, disability, class, but covers all of discrimination, and how they "intersect" to in the creation and sustaining of systems of oppression (C.I.J, 2022). I believe, establishing a greater emphasis on equity, intersectionality and social good in the empathy phase of the XAI design framework, along with meaningful metrics to determine success in the test stage are important areas of research that would be highly beneficial to designers, organizations, and society as a whole

References

Abadi, Mart&#x27;in, Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., … others. (2016). Tensorflow: A system for large-scale machine learning. In 12th $USENIX$ Symposium on Operating Systems Design and Implementation ($OSDI$ 16) (pp. 265–283).

Adadi, A., & Berrada, M. (2018). Peeking inside the black box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

Akash, K., Polson, K., Reid, T., & Jain, N. (2019). Improving human-machine collaboration through transparency-based feedback–part I: Human trust and workload model. IFAC-PapersOnLine, 51(34), 315-321.

Alós-Ferrer, C., & Farolfi, F. (2019). Trust games and beyond. Frontiers in neuroscience, 13, 887.

American Psychological Association, American Educational Research Association, & National Council on Measurement in Education. (1954). Technical recommendations for psychological tests and diagnostic techniques (Vol. 51, No. 2). American Psychological Association.

Ancker, J. S., Edwards, A., Nosal, S., Hauser, D., Mauer, E., & Kaushal, R. (2017). Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. BMC medical informatics and decision making, 17(1), 1-9.

Annett, J. (2002). A note on the validity and reliability of ergonomics methods.

Ashraf, B. N. (2020). Economic impact of government interventions during the COVID-19 pandemic: International evidence from financial markets. Journal of behavioral and experimental finance, 27, 100371.

Balfe, N., Sharples, S., & Wilson, J. R. (2015). Impact of automation: Measurement of performance, workload and behavior in a complex control environment. Applied Ergonomics, 47, 52–64. https://doi.org/10.1016/j.apergo.2014.08.002

Barsuk, J. H., Cohen, E. R., Feinglass, J., McGaghie, W. C., & Wayne, D. B. (2009). Use of simulation-based education to reduce catheter-related bloodstream infections. Archives of internal medicine, 169(15), 1420-1423.

Barsuk, J. H., Cohen, E. R., McGaghie, W. C., & Wayne, D. B. (2010). Long-term retention of central venous catheter insertion skills after simulation-based mastery learning. Academic Medicine, 85(10), S9-S12.

Bauer, G. R., & Lizotte, D. J. (2021). Artificial intelligence, intersectionality, and the future of public health. American Journal of Public Health, 111(1), 98-100.

Bauhs, J. A., & Cooke, N. J. (1994, April). Is knowing more really better? Effects of system development information in human-expert system interactions. In Conference Companion on Human Factors in Computing Systems (pp. 99-100).

Bayer, S., Gimpel, H., & Markgraf, M. (2021). The role of domain expertise in trusting and following explainable AI decision support systems. Journal of Decision Systems, 1-29.

Belkin, N. J. (1984). Cognitive models and information transfer. Social Science Information Studies, 4(2-3), 111-129.

Bernard, J., Hutter, M., Sedlmair, M., Zeppelzauer, M., & Munzner, T. (2021). A Taxonomy of Property Measures to Unify Active Learning and Human-centered Approaches to Data Labeling. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-42.

Bhatt, U., Zhang, Y., Antorán, J., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G. G., Krishnan, R., Stanley, J., Tickoo, O., Nachman, L., Chunara, R., Weller, A., & Xiang, A. (2020). Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. arXiv. arXiv:2011.07586

Bradski, G., & Kaehler, A. (2008). Learning OpenCV: Computer vision with the OpenCV library. " O&#x27;Reilly Media, Inc."

Brooke, J. (1986). System usability scale (SUS): a quick-and-dirty method of system evaluation user information. Reading, UK: Digital equipment co ltd, 43, 1-7.Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. Army research lab Aberdeen proving ground md human research and engineering directorate.

Cascio, W. F., & Aguinis, H. (2001). The federal uniform guidelines on employee selection procedures (1978) an update on selected issues. Review of Public Personnel Administration, 21(3), 200-218.

Center for Intersectional Justice C.I.J (2022). What is Intersectionality?. Retrieved . 2022, October 28 2022, October 28 https://www.intersectionaljustice.org/what-is-intersectionality

Chen, J. Y., & Barnes, M. J. (2013). Human-agent teaming for multi-robot control: A literature review ARL-TR-6328 Army Research Laboratory.

Chen, J. Y., Procci, K., Boyce, M., Wright, J., Garcia, A., & Barnes, M. (2014). Situation awareness-based agent transparency. Army research lab Aberdeen proving ground md human research and engineering directorate.

Chollet, F., & others. (2015). Keras. GitHub. Retrieved from https://github.com/fchollet/keras

Ciston, S. (2019). Intersectional AI is essential: polyvocal, multimodal, experimental methods to save artificial intelligence. Journal of Science and Technology of the Arts, 11(2), 3-8. https://doi.org/10.7559/citarj.v11i2.665

Cooper, R., Fox, J., Farringdon, J., & Shallice, T. (1996). A systematic methodology for cognitive modelling. Artificial Intelligence, 85(1-2), 3-44.

Cornelissen, M., McClure, R., Salmon, P. M., & Stanton, N. A. (2014). Validating the strategies analysis diagram: assessing the reliability and validity of a formative method. Applied ergonomics, 45(6), 1484-1494.

Craig, C., Klein, M., Griswold, J., Gaitonde, K., McGill, T. & Halldorsson, A. (2012). Using cognitive task analysis to identify critical decisions in laparoscopic environments. Human factors, 54(6), 1025-1039

Crenshaw, K. (1989). 1 995. Mapping the margins: Intersectionality, identity politics, and violence against women of color. Critical race theory: The writings that formed the movement, ed. K. Crenshaw, N. Gotanda, G. Peller, and K. Thomas, 357-83.

Crichton, M. T., Flin, R., & Rattray, W. A. (2000). Training decision makers–tactical decision games. Journal of contingencies and crisis management, 8(4), 208-217.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. Psychological bulletin, 52(4), 281.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. American Institute of Aeronautics and Astronautics, 2, 557–562. https://doi.org/10.2514/6.2004-6313

Cummings, M. L. (2017). Automation bias in intelligent time critical decision support systems. In Decision Making in Aviation (pp. 289-294). Routledge.

Danks, D., & London, A. J. (2017, August). Algorithmic Bias in Autonomous Systems. In IJCAI (Vol. 17, pp. 4691-4697).

Datta, A., Tschantz, M. C., & Datta, A. (2015). Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination. Proceedings on Privacy Enhancing Technologies, 1, 92-112. https://doi.org/10.1515/popets-2015-0007

David, N. (2013). Validating simulations. In Simulating social complexity (pp. 135-171). Springer, Berlin, Heidelberg.

Davies, D., & Dodd, J. (2002). Qualitative research and the question of rigor. Qualitative health research, 12(2), 279-289.

de Fine Licht, J. (2011). Do we really want to know? The potentially negative effect of transparency in decision making on perceived legitimacy. Scandinavian Political Studies, 34(3), 183–201. https://doi.org/10.1111/j.1467-9477.2011.00268.x

de Fine Licht, K., & de Fine Licht, J. (2020). Artificial intelligence, transparency, and public decision-making. AI & Society, 35, 917–926. https://doi.org/10.1007/s00146-020-00960

de Vries, P., Midden, C., & Bouwhuis, D. (2003). The effects of errors on system trust, self-confidence, and the allocation of control in route planning. International Journal of Human Computer Studies, 58(6), 719–735. https://doi.org/10.1016/S1071-5819(03)00039-9

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). Ieee

DeWine, M., Husted, J. & Acton, A. (2020). Director's Stay Safe Ohio Order. Ohio Department of Health. https://coronavirus.ohio.gov/static/publicorders/Directors-Stay-Safe-Ohio-Order.pdf

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. arXiv preprint arXiv:1710.00794.

Dorst, K. (2011). The core of 'design thinking' and its application. Design Studies, 32(6), 521–532. https://doi.org/10.1016/j.destud.2011.07.006

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. International Journal of Human Computer Studies, 58(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. Human Factors, 44(1), 79–94. https://doi.org/10.1518/0018720024494856

Embretson, S. E. (1983). Construct validity: Construct representation versus nomothetic span. Psychological Bulletin, 93, 179–197

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. Human Factors, 37(1), 32–64. https://doi.org/10.1518/001872095779049543

European Union. (2018). Coordinated Plan on Artificial Intelligence, “made in Europe”, 7.12.2018 COM(2018) 795 2018, European Union.

Ezer, N., Bruni, S., Cai, Y., Hepenstal, S. J., Miller, C. A., & Schmorrow, D. D. (2019, November). Trust engineering for human-AI teams. In Proceedings of the Human Factors and

Fan, X., & Yen, J. (2010). Modeling cognitive loads for evolving shared mental models in human–agent collaboration. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 41(2), 354–367. https://doi.org/10.1109/TSMCB.2010.2053705

Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. Big Data & Society, 6(1), 1–14. https://doi.org/10.1177/2053951719860542

Foulds, J. R., Islam, R., Keya, K. N., & Pan, S. (2020, April). An intersectional definition of fairness. In 2020 IEEE 36th International Conference on Data Engineering (ICDE) (pp. 1918-1921). IEEE.

Garcia, M. (2016). Racist in the machine: The disturbing implications of algorithmic bias. World Policy Journal, 33(4), 111-117.

Gasparini, A. (2015). Perspective and use of empathy in design thinking. ACHI, the eight international conference on advances in computer-human interactions. 2015.

Gestwicki, P., & McNely, B. (2012). A case study of a five-step design thinking process in educational museum game design. Proceedings of Meaningful Play, USA, 1–30.

Gilbert, J. (2021, May). Equitable AI. In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-2).

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goddard, K., Roudsari, A., & Wyatt, J. C. (2012). Automation bias: A systematic review of frequency, effect mediators, and mitigators. Journal of the American Medical Informatics Association, 19(1), 121–127. https://doi.org/10.1136/amiajnl-2011-000089

Goldmann, D. A., & Pier, G. B. (1993). Pathogenesis of infections related to intravascular catheterization. Clinical microbiology reviews, 6(2), 176-192.

Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1, No. 2). Cambridge: MIT press.

Goodman, B., & Flaxman, S. (2016). European Union regulations on algorithmic decision making and a “right to explanation.” Presented at the ICML Workshop on Human Interpretability in Machine Learning, New York, NY.

Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. Journal of Field Robotics, 37(3), 362-386.

Gunning, D., & Aha, D. (2019). DARPA’s explainable artificial intelligence (XAI) program. AI magazine, 40(2), 44-58.

Gusenleitner, N., Siedl, S., Stübl, G., Polleres, A., Recski, G., Sommer, R., ... & Moser, B. A. (2019). Facing mental workload in AI-transformed working environments. In H-Workload 2019: 3rd International Symposium on Human Mental Workload: Models and Applications (Works in Progress) (p. 91).

Ha, C. H., Kim, J. H., Lee, S. J., & Seong, P. H. (2006). Investigation on relationship between information flow rate and mental workload of accident diagnosis tasks in NPPs. IEEE Transactions on Nuclear Science, 53(3), 1450-1459.

Haibe-Kains, B., Adam, G. A., Hosny, A., Khodakarami, F., Waldron, L., Wang, B., ... & Aerts, H. J. (2020). Transparency and reproducibility in artificial intelligence. Nature, 586(7829), E14-E16.

Hale, T., Petherick, A., Phillips, T., & Webster, S. (2020). Variation in government responses to COVID-19. Blavatnik school of government working paper, 31.

Haleem, A., Javaid, M., Vaishya, R., & Deshmukh, S. G. (2020). Areas of academic research with the impact of COVID-19. The American Journal of Emergency Medicine.

Halgren, S. L. (1993). In search of optimal human-expert system explanations: Empirical studies of human-human and human-expert system interactions. Rice University.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. Human Factors, 53(5), 517–527. https://doi.org/10.1177/0018720811417254

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183). North-Holland.

Haselton, M. G., Nettle, D., & Murray, D. R. (2015). The evolution of cognitive bias. In D. M. Buss (Ed.), The handbook of evolutionary psychology (2nd ed., pp. 968–987). John Wiley & Sons, Inc. https://doi.org/10.1002/9781119125563.evpsych241

Hayes-Roth, B., & Hayes-Roth, F. (1979). A cognitive model of planning. Cognitive science, 3(4), 275-310.

Helldin, T. (2014). Transparency for Future Semi-Automated Systems: Effects of transparency on operator performance, workload and trust [Doctoral dissertation, Örebro Universitet]. DiVA. https://www.diva-portal.org/smash/get/diva2:710832/FULLTEXT02

Helldin, T., Falkman, G., Riveiro, M., & Davidsson, S. (2013, October). Presenting system uncertainty in automotive UIs for supporting trust calibration in autonomous driving. In Proceedings of the 5th international conference on automotive user interfaces and interactive vehicular applications (pp. 210-217).

Hengstler, M., Enkel, E., & Duelli, S. (2016). Applied artificial intelligence and trust–The case of autonomous vehicles and medical assistance devices. Technological Forecasting and Social Change, 105, 105–120. https://doi.org/10.1016/j.techfore.2015.12.014

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. Human Factors, 57(3), 407–434. https://doi.org/10.1177/0018720814547570

Hoffman, R. R., Feltovich, P. J., Ford, K. M., & Woods, D. D. (2002). A rose by any other name... would probably be given an acronym [cognitive systems engineering]. IEEE Intelligent Systems, 17(4), 72–80. https://doi.org/10.1109/MIS.2002.1024755

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects. arXiv preprint arXiv:1812.04608.

Hollnagel, E., & Woods, D. D. (1983). Cognitive systems engineering: New wine in new bottles. International Journal of Man-Machine Studies, 18(6), 583–600. https://doi.org/10.1016/S0020-7373(83)80034-0

Horvitz, E. J., & Barry, M. (2013). Display of information for time-critical decision making. arXiv. arXiv:1302.4959.

IBM (2021). What Is Machine Learning? Website: https://www.ibm.com/topics/machine-learning, (retrieved 12/12/2022)

Ikuma, L. H., Harvey, C., Taylor, C. F., & Handal, C. (2014). A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. Journal of Loss Prevention in the Process Industries, 32, 454–465. https://doi.org/10.1016/j.jlp.2014.11.001

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. Business Horizons, 61(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). The measurement of the propensity to trust automation. In J. Y. C. Chen, & G. Fragomeni (Eds.), Lecture notes in computer science: Vol. 11575. Virtual, augmented and mixed reality: Applications and case studies (pp. 476–489). Springer. https://doi.org/10.1007/978-3-030-21565-1_32

Jian, J., Bisantz, A. M., Drury, C. G.: Foundations for an empirically determined scale of trust in automated systems. International Journal of Cognitive Ergonomics, 4(1), 53-71 (2000).

Johari, A., & Swami, P. D. (2020, February). Comparison of autonomy and study of deep learning tools for object detection in autonomous self driving vehicles. In 2nd International Conference on Data, Engineering and Applications (IDEA) (pp. 1-6). IEEE.

Johnson, G. M. (2020). Algorithmic bias: On the implicit biases of social technology. Synthese, 1–21. https://doi.org/10.1007/s11229-020-02696-y

Jones, D. G., & Endsley, M. R. (1996). Sources of situation awareness errors in aviation. Aviation, Space, and Environmental Medicine, 67(6), 507–512.

Jones, G. R., & George, J. M. (1998). The experience and evolution of trust: Implications for cooperation and teamwork. Academy of Management Review, 23(3), 531–546. https://doi.org/10.5465/AMR.1998.926625

Kane, M. (2006). Validation. In R. Brennan (Ed.), Educational measurement (4th ed., pp. 17-64), Westport, CT: American Council on Education and Praeger

Kane, M. (2013). The argument-based approach to validation. School Psychology Review, 42(4), 448-457.

Karnon, J. (2003). Alternative decision modeling techniques for the evaluation of health care technologies: Markov processes versus discrete event simulation. Health Economics, 12(10), 837–848. https://doi.org/10.1002/hec.770

Keehner, M., Gorin, J. S., Feng, G., & Katz, I. R. (2017). Developing and validating cognitive models in assessment. The handbook of cognition and assessment: Frameworks, methodologies, and applications, 75-101.

Kembel, G. (2009). Awakening creativity. https://programarchive.chq.org/ci/sessions/6409/view accessed April, 2021 (August 14, 2009 Presentation at the Chautauqua Institution). Chautauqua Institution.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M., & Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. arXiv preprint arXiv:1609.04836.

Kim, B., Park, J., & Suh, J. (2020). Transparency and accountability in AI decision support: Explaining and visualizing convolutional neural networks for text information. Decision Support Systems, 134, Article 113302. https://doi.org/10.1016/j.dss.2020.113302

King, S. B., Babb, J. D., Bates, E. R., Crawford, M. H., Dangas, G. D., Voeltz, M. D., & White, C. J. (2015). COCATS 4 Task Force 10: training in cardiac catheterization. Journal of the American College of Cardiology, 65(17), 1844-1853.

Klein, G. A. (1993). A recognition-primed decision (RPD) model of rapid decision making. Decision making in action: Models and methods, 5(4), 138-147.

Klein, G. A. (1999). Sources of power: How people make decisions. MIT press.

Kratsios, M. (2019). The National Artificial Intelligence Research and Development Strategic Plan: Update. 50. National Science and Technology Council.

Kumar, V. (2012). 101 design methods: A structured approach for driving innovation in your organization. John Wiley & Sons.

Landry, M., Malouin, J. L., & Oral, M. (1983). Model validation in operations research. European Journal of Operational Research, 14(3), 207-220.

Larsson, S., & Heintz, F. (2020). Transparency in artificial intelligence. Internet Policy Review, 9(2), 1–16. https://doi.org/10.14763/2020.2.1469

Law, A. W. D. Kelton. (1991). Simulation Modeling and Analysis.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, J. D., & Moray, N. (1994). Trust, self-confidence, and operators' adaptation to automation. International Journal of Human-Computer Studies, 40(1), 153–184. https://doi.org/10.1006/ijhc.1994.1007

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. Human Factors, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lewandowsky, S., Mundy, M., & Tan, G. (2000). The dynamics of trust: comparing humans to automation. Journal of Experimental Psychology: Applied, 6(2), 104.

Lewicki, R. J., Tomlinson, E. C., & Gillespie, N. (2006). Models of interpersonal trust development: Theoretical approaches, empirical evidence, and future directions. Journal of Management, 32(6), 991–1022. https://doi.org/10.1177/0149206306294405

Lewis, C., Polson, P. G., Wharton, C., & Rieman, J. (1990, March). Evaluating a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 235-242).

Lewis, J. R., & Sauro, J. (2009, July). The factor structure of the system usability scale. In International conference on human centered design (pp. 94-103). Springer, Berlin, Heidelberg.

Lewis, M., Sycara, K., & Walker, P. (2018). The role of trust in human-robot interaction. In H. A. Abbass, J. Scholz, & D. J. Reid (Eds.), Foundations of trusted autonomy (pp. 135–159). Springer. https://doi.org/10.1007/978-3-319-64816-3_8

Lintern, G. (2010). A comparison of the decision ladder and the recognition-primed decision model. Journal of Cognitive Engineering and Decision Making, 4(4), 304-327.

López-Tapia, S., Molina, R., & de la Blanca, N. P. (2018). Using machine learning to detect and localize concealed objects in passive millimeter-wave images. Engineering Applications of Artificial Intelligence, 67, 81-90.

Lu, Y., & Sarter, N. (2019). Eye tracking: a process-oriented method for inferring trust in automation as a function of priming and system reliability. IEEE Transactions on Human-Machine Systems, 49(6), 560-568.

Lucas, F. (2018). Techniques for Empathy Interviews in Design Thinking. Web Design Envato Tuts. https://webdesign.tutsplus.com/articles/techniques-of-empathy-interviews-in-design-thinking--cms-31219

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.

Lyons J. B., & Havig P. R. (2014). Transparency in a human-machine context: Approaches for fostering shared awareness/intent. In R. Shumaker, & S. Lackey (Eds.), Lecture notes in computer science: Vol. 8525. Virtual, augmented and mixed reality: Designing and developing virtual and augmented environments (pp. 181–190). Springer. https://doi.org/10.1007/978-3-319-07458-0_18

Lyons, J. B., Nam, C. S., Jessup, S. A., Vo, T. Q., & Wynne, K. T. (2020). The role of individual differences as predictors of trust in autonomous security robots. IEEE International Conference on Human-Machine Systems (ICHMS), 1–5. https://doi.org/10.1109/ichms49158.2020.9209544

Lyons, J. B., Sadler, G. G., Koltai, K., Battiste, H., Ho, N. T., Hoffmann, L. C., Smith, D., Johnson, W., & Shively, R. (2017). Shaping trust through transparent design: Theoretical and experimental guidelines. Advances in Intelligent Systems and Computing, 499, 127–136. https://doi.org/10.1007/978-3-319-41959-6_11

Madhav, A. V., & Tyagi, A. K. (2023). Explainable Artificial Intelligence (XAI): Connecting Artificial Decision-Making and Human Trust in Autonomous Vehicles. In Proceedings of Third International Conference on Computing, Communications, and Cyber-Security (pp. 123-136). Springer, Singapore.

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. Theoretical Issues in Ergonomics Science, 8(4), 277-301.

Malle, B. F., & Ullman, D. (2021). A multidimensional conception and measure of human-robot trust. In C. S. Nam & J. B. Lyons (Eds.), Trust in human-robot interaction (pp. 3–25). Elsevier. https://doi.org/10.1016/b978-0-12-819472-0.00001-0

Mann, N. C. (2016). National Emergency Medical Services Information System (NEMSIS). Prehospital Emergency Care, 10(3), 314–316. https://nemsis.org/

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. The Academy of Management Review, 20(3), 709–734. https://doi.org/10.2307/258792

McCarthy, J. (2007). What is artificial intelligence. URL: http://www-formal. stanford. edu/jmc/whatisai. html.

McGuirl, J. M., & Sarter, N. B. (2006). Supporting trust calibration and the effective use of decision aids by presenting dynamic system confidence information. Human factors, 48(4), 656-665..

Mercado, J. E., Rupp, M. A., Chen, J. Y. C., Barnes, M. J., Barber, D., & Procci, K. (2016). Intelligent agent transparency in human-agent teaming for multi-UxV management. Human Factors, 58(3), 401–415. https://doi.org/10.1177/0018720815621206

Merritt, S. M., Unnerstall, J. L., Lee, D., & Huber, K. (2015). Measuring individual differences in the perfect automation schema. Human Factors, 57(5), 740–753. https://doi.org/10.1177/0018720815581247

Middleton, F. (2019). The 4 Types of Validity. Website, ReScribbr.com (retrieved 12/12/2020) https://www.scribbr.com/methodology/types-of-validity/

Militello, L. G., & Hutton, R. J. (1998). Applied cognitive task analysis (ACTA): a practitioner's toolkit for understanding cognitive task demands. Ergonomics, 41(11), 1618-1641.

Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 11(3-4), 1-45.

Mosier, K. L., & Skitka, L. J. (1996). Human decision makers and automated decision aids: Made for each other?. In M. Mouloua, & R. Parasuraman (Eds.), Automation and human performance: Theory and applications (pp. 201–220). Taylor & Francis.

Muir, B. M. (1987). Trust between humans and machines, and the design of decision aids. International Journal of Man-Machine Studies, 27(5–6), 527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Mundhenk, T. N., Chen, B. Y., & Friedland, G. (2019). Efficient saliency maps for explainable ai. arXiv preprint arXiv:1911.11293.

Naiseh, M., Al-Thani, D., Jiang, N., & Ali, R. (2022). How the different explanation classes impact trust calibration: The case of clinical decision support systems. International Journal of Human-Computer Studies, 102941.

Okamura, K., & Yamada, S. (2020). Adaptive trust calibration for human-AI collaboration. Plos one, 15(2), e0229132.

Palacio, S., Lucieri, A., Munir, M., Ahmed, S., Hees, J., & Dengel, A. (2021). Xai handbook: Towards a unified framework for explainable ai. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 3766-3775).

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. Human Factors, 39(2), 230–253. https://doi.org/10.1518/001872097778543886

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model of types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics — Part A: 30(3), 286–297. 10.1109/3468.844354

Pham, H., Dai, Z., Xie, Q., & Le, Q. V. (2021). Meta pseudo labels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 11557-11568).

Poundstone, W. (1993). Prisoner's Dilemma/John Von Neumann, Game theory and the puzzle of the bomb. Anchor.

Pronin, E., Lin, D. Y., & Ross, L. (2002). The bias blind spot: Perceptions of bias in self versus others. Personality and Social Psychology Bulletin, 28(3), 369-381.

Rasmussen, J. (1974). The human data processor as a system component. Bits and pieces of a model. Report No. Risø -M-1722.

Riga, C. V., Bicknell, C. D., Hamady, M. S., & Cheshire, N. J. (2011). Evaluation of robotic endovascular catheters for arch vessel cannulation. Journal of vascular surgery, 54(3), 799-809.

Rosebrock A. (2019) Fine-tuning with Keras and Deep Learning. PyImageSearch. https://www.pyimagesearch.com/2019/06/03/fine-tuning-with-keras-and-deep-learning/

Rupp, A. A., & Leighton, J. P. (Eds.). (2016).The Wiley handbook of cognition and assessment: Frameworks, methodologies, and applications. John Wiley & Sons.

Sarter, N. B., Woods, D. D., & Billings, C. E. (1997). Automation surprises. In G. Salvendy (Ed.), Handbook of human factors & ergonomics (2nd ed., pp. 1926–1943). Wiley.

Schaefer, K. E., Chen, J. Y. C., Szalma, J. L., & Hancock, P. A. (2016). A Meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors, 58(3), 377–400. https://doi.org/10.1177/0018720816634228

Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. Journal of Decision Systems, 29(4), 260-278.

Schulz, C. M., Krautheim, V., Hackemann, A., Kreuzer, M., Kochs, E. F., & Wagner, K. J. (2016). Situation awareness errors in anesthesia and critical care in 200 cases of a critical incident reporting system. BMC Anesthesiology, 16(1), 4–10. https://doi.org/10.1186/s12871-016-0172-7

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Sheridan, T. B. (1997). Task analysis, task allocation and supervisory control. In M. G. Helander, T. K. Landauer, & P. V. Prabhu (Eds.), Handbook of human-computer interaction (2nd ed., pp. 87–105). North-Holland. https://doi.org/10.1016/B978-044481862-1.50071-6

Sheridan, T. B., & Verplank, W. L. (1978). Human and computer control of undersea teleoperators. Massachusetts Institute of Technology. https://apps.dtic.mil/sti/citations/ADA057655

Silva, C., Vieira, J., Campos, J. C., Couto, R., & Ribeiro, A. N. (2020). Development and Validation of a Descriptive Cognitive Model for Predicting Usability Issues in a Low-Code Development Platform. Human Factors, 0018720820920429.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2013). Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:1312.6034.

Simpson, J. A. (2007). Foundations of interpersonal trust. In A. W. Kruglanski, & E. T. Higgins (Eds.), Social psychology: Handbook of basic principles (pp. 587–607). The Guilford Press.

Smith, T. J., Henning, R. A., Wade, M. G., & Fisher, T. (2014). Variability in human performance. CRC Press.

Soh, B. K. (2007). Validation of the recognition-primed decision model and the roles of common-sense strategies in an adversarial environment (Doctoral dissertation, Virginia Tech).

Stanton, N. A. (2016). On the reliability and validity of, and training in, ergonomics methods: a challenge revisited. Theoretical Issues in Ergonomics Science, 17(4), 345-353.

Stanton, N. A., & Baber, C. (2005). Validating task analysis for error identification: reliability and validity of a human error prediction technique. Ergonomics, 48(9), 1097-1113.

Stanton, N. A., & Stevenage, S. V. (1998). Learning to predict human error: issues of acceptability, reliability and validity. Ergonomics, 41(11), 1737-1756.

Stanton, N. A., & Young, M. S. (1999). What price ergonomics?. Nature, 399(6733), 197-198.

Stent: Purpose, Procedure, and Risks. (2017). Retrieved 19 April 2020, from https://www.healthline.com/health/stent#procedure

Stone, P. B. (2019). Agent-based simulation of artificial-intelligence-assisted transfer of care [Master’s Thesis, Wright State University]. Core Scholar. https://corescholar.libraries.wright.edu/cgi/viewcontent.cgi?article=3281&context=etd_all

Stone, P. B. (2020) Improving Decision Support Systems Through Context and Demand Aware Augmented Intelligence in Dynamic Joint Cognitive Systems. Unpublished manuscript.

Stone, P. B., Nelson, H. M., Fendley, M. E., & Ganapathy, S. (2021). Development of a novel hybrid cognitive model validation framework for implementation under COVID-19 restrictions. Human Factors and Ergonomics in Manufacturing & Service Industries, 31(4), 360-374.

Stone, P.B., Jessup, S.A., Ganapathy, S, Harel, A, (2022) Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support. International Journal of Human Computer Interaction

Strobel, M. (2019). Aspects of transparency in machine learning. Proceedings of the International Conference on Autonomous Agents and MultiAgent Systems, Canada, 2449–2451. https://doi.org/10.5555/3306127.3332143

Strube, G. "Cognitive modeling: research logic in cognitive science." (2001): 2124-2128.

Sy, M., O'Leary, N., Nagraj, S., El-Awaisi, A., O'Carroll, V., & Xyrichis, A. (2020). Doing interprofessional research in the COVID-19 era: a discussion paper. Journal of Interprofessional Care, 34(5), 600-606.

Teh, E., Jamson, S., Carsten, O., & Jamson, H. (2014). Temporal fluctuations in driving demand: The effect of traffic complexity on subjective measures of workload and driving performance. Transportation research part F: traffic psychology and behavior, 22, 207-217.

Thelisson, E., Padh, K., & Celis, L. E. (2017). Regulatory mechanisms and algorithms towards trust in AI/ML. Proceedings of the International Joint Conference on Artificial Intelligence, Australia, 1–5.

Thoring, K., & Müller, R. M. (2011). Understanding the creative mechanisms of design thinking: an evolutionary approach. Proceedings of the Second Conference on Creativity and Innovation in Design, Netherlands, 137–147. https://doi.org/10.1145/2079216.2079236

Thoroman, B., Salmon, P., & Goode, N. (2019). Evaluation of construct and criterion-referenced validity of a systems-thinking based near miss reporting form. Ergonomics, 63(2), 210-224.

Tomar, S. (2006). Converting video formats with FFmpeg. Linux Journal, 2006(146), 10.

Topol, E. (2019). Deep medicine: How artificial intelligence can make healthcare human again. Basic Books.

Tseng, K. K., Zhang, R., Chen, C. M., & Hassan, M. M. (2021). DNetUnet: a semi-supervised CNN of medical image segmentation for super-computing AI service. The Journal of Supercomputing, 77(4), 3594-3615.

Turek, M. (2020). Explainable Artificial Intelligence (XAI). 2016. URL: https://www. Dapra. mil/program/explainable-artificial-intelligence [cited July, 2017].

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. Science, 185(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Vinod, A. P., Tang, Y., Oishi, M. M., Sycara, K., Lebiere, C., & Lewis, M. (2016). Validation of cognitive models for collaborative hybrid systems with discrete human input. In 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 3339-3346). IEEE.

Vitulano, S., Ruberto, C. D., & Nappi, M. (1995). AI based image segmentation. In International Conference on Image Analysis and Processing (pp. 429-434). Springer, Berlin, Heidelberg.

Wachter, S., Mittelstadt, B., & Russell, C. (2017). Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. http://dx.doi.org/10.2139/ssrn.3063289

Wagenaar, W. A., Hudson, P. T., & Reason, J. T. (1990). Cognitive failures and accidents. Applied Cognitive Psychology, 4(4), 273-294

Wang, D., Khosla, A., Gargeya, R., Irshad, H., & Beck, A. H. (2016). Deep learning for identifying metastatic breast cancer. arXiv preprint arXiv:1606.05718.

Wickens, C. D., & Dixon, S. R. (2007). The benefits of imperfect diagnostic automation: A synthesis of the literature. Theoretical Issues in Ergonomics Science, 8(3), 201-212.

Woods, D. D. (1985). Cognitive technologies: The design of joint human-machine cognitive systems. AI Magazine, 6(4), 86–92. https://doi.org/10.1609/aimag.v6i4.511

World Health Organization (2020). COVID-19 Weekly Epidemiological Update. Retrieved November 11 from https://www.who.int/publications/m/item/weekly-epidemiological-update---10-november-2020.

Wu, C., Cha, J., Sulek, J., Zhou, T., Sundaram, C. P., Wachs, J., & Yu, D. (2020). Eye-tracking metrics predict perceived workload in robotic surgical skills training. Human factors, 62(8), 1365-1386

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019, October). Explainable AI: A brief survey on history, research areas, approaches and challenges. In CCF international conference on natural language processing and Chinese computing (pp. 563-574). Springer, Cham.

Yoko, K. (2006). Student pilot situational awareness: The effects of trust in technology (Doctoral dissertation, Embry-Riddle Aeronautical University).

Yuan, L., Chen, D., Chen, Y. L., Codella, N., Dai, X., Gao, J., ... & Zhang, P. (2021). Florence: A new foundation model for computer vision. arXiv preprint arXiv:2111.11432.

Zhang, Y., Liao, Q. V., & Bellamy, R. K. (2020, January). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 295-305)

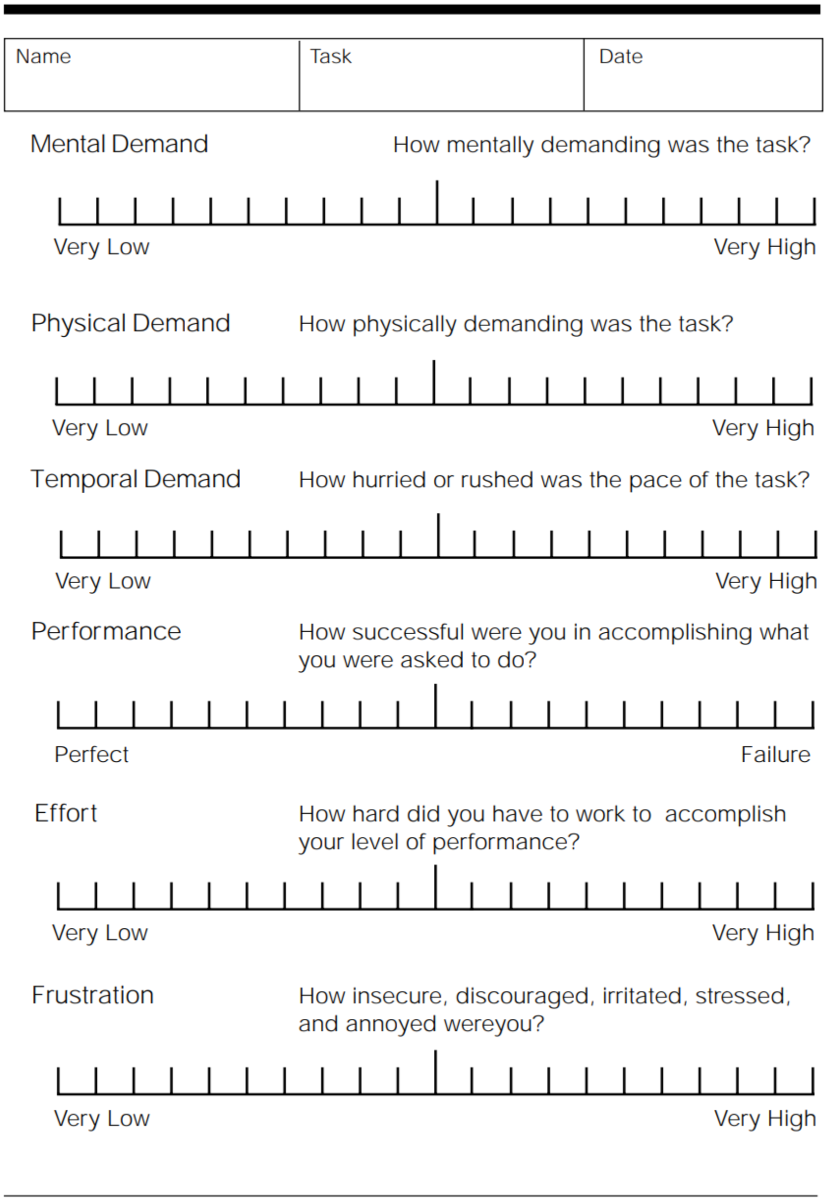
Zhang, Z. (2019, October). Technologies raise the effectiveness of airport security control. In 2019 IEEE 1st International Conference on Civil Aviation Safety and Information Technology (ICCASIT) (pp. 431-434). IEEE.

Zhao, W., Shen, L., Islam, M. T., Qin, W., Zhang, Z., Liang, X., ... & Li, X. (2021). Artificial intelligence in image-guided radiotherapy: a review of treatment target localization. Quantitative Imaging in Medicine and Surgery, 11(12), 4881.

Zuk, T., & Carpendale, S. (2007, June). Visualization of uncertainty and reasoning. In A. Butz, B. Fisher, A. Krüger, P. Olivier, & S. Owada (Eds.), Lecture notes in computer science: Vol. 4569. Smart graphics (pp. 164-177). Springer. https://doi.org/10.1007/978-3-540-73214-3_15

APPENDIX A -NASA TASK LOAD INDEX

Hart and Staveland’s NASA Task Load Index (TLX) method assesses workload on five 7-point scales. Increments of high, medium, and low estimates for each point result in 21 gradations on



Name	Task	Date

Mental Demand How mentally demanding was the task?

Very Low Very High

Physical Demand How physically demanding was the task?

Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?

Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?

Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?

Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed wereyou?

Very Low Very High

APPENDIX B – TRUST IN AUTOMATION

The Adapted Trust in Automated Systems (Jian, Bisantz, & Drury, 2000), is a scale for measuring intention to trust in users.

Instructions: Below is a list of statements for evaluating trust. There are several items for you to rate intensity of your feeling of trust, or your impression of the automation while engaging in a task. Please select the option which describes your feeling or your impression using the 7-point scale ranging from 1 (not at all) to 7 (extremely).

1. The automation is deceptive. (R)
2. The automation behaves in an underhanded manner. (R)
3. I am wary of the automation. (R)
4. The automation’s actions will have a harmful or injurious outcome. (R)
5. I am confident in the automation.
6. The automation provides security.
7. The automation has integrity.
8. The automation is dependable.
9. The automation is reliable.
10. I can trust the automation.
11. I am familiar with the automation

APPENDIX C – PROPENSITY TO TRUST

The Adapted Propensity to Trust in Technology (Schneider et al., 2017), measures a user's baseline level of propensity to trust or risk aversion.

Instructions: For the below listed items, please read each statement carefully. Using the 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree), select the answer that most accurately describes your feelings.

1. Generally, I trust automated agents.
2. Automated agents help me solve many problems.
3. I think it's a good idea to rely on automated agents for help.
4. I don't trust the information I get from automated agents. (R)
5. Automated agents are reliable.
6. I rely on automated agents.

APPENDIX D - SYSTEM USABILITY SCALE

The System Usability Scale (SUS), Brooke (1986) is used to determine the usability of a system. Participants are asked to score the following 10 items with one of five responses that range from Strongly Agree to Strongly disagree:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

VITA

Paul Benjamin Stone

1999	B.Eng Engineering Design: Mechanical, University of Huddersfield
1999-2001	Design Engineer, Hygood-Macron
2001-2009	Research Scientist, Defence Science and Technology Laboratory
2009-2017	Senior Research Scientist, Defence Science and Technology Laboratory
2018-2022	Graduate Research/Teaching Assistant, Wright State University
2019	M.S. Industrial and Human Factors Engineering, Wright State University
2022	PhD. Industrial and Human Factors Engineering, Wright State University

Field of Study

Industrial and Human Factors Engineering

Publications

Stone, P.B., Jessup, S.A., Ganapathy, S, Harel, A, (2022) Design Thinking Framework for Integration of Transparency Measures in Time-Critical Decision Support. International Journal of Human Computer Interaction

Stone, P. B., Nelson, H. M., Fendley, M. E., & Ganapathy, S. (2021). Development of a novel hybrid cognitive model validation framework for implementation under COVID-19 restrictions. Human Factors and Ergonomics in Manufacturing & Service Industries, 31(4), 360-374.

Stone, P. B., Ganapathy, S., Fendley, M.E., Akilan, L., (2021). Integrating Wearable Devices in Real-Time Computer Applications of Petrochemical Systems, International Human Machine Interaction and Collaboration Conference. World Academy of Science, Engineering and Technology

Stone P.B., Ganapathy S. (2019). Agent-Based Simulation of Artificial-Intelligence-Assisted Transfer of Care Proceedings of the 2019 IISE Annual Conference.

Stone, P. B. (2019). Agent-based simulation of artificial-intelligence-assisted transfer of care [Master's Thesis, Wright State University]. Core Scholar.

Presentations

Agent-Based Simulation of Artificial-Intelligence-Assisted, Transfer of Care. IISE Annual Conference (May 2019, Orlando, Florida).

Integrating Wearable Devices in Real-Time Computer Applications of Petrochemical Systems. International Human Machine Interaction and Collaboration Conference. (September 2021, San Francisco, California – Virtual Conference).

Understanding the phenomenology of rotorblade scintillation (Halo effect) and potential mitigations. Defence Science and Technology Laboratory Annual conference (August 2012, Richmond, UK)