USAARL-TECH-CR--2022-22

Methods and Measures to Evaluate Technologies that Influence Aviator Decision Making and Situation Awareness

Emilie Roth, Devorah Klein, Christen Sushereba, Katie Ernst,
& Lauren Militello

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
31-03-2022	Contract Report	

4. TITLE AND SUBTITLE

Methods and Measures to Evaluate Technologies that Influence Aviator Decision Making and Situation Awareness

5a. CONTRACT NUMBER

Prime contract No. 80ARC020D006

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Roth, E.[1], Klein, D.[2], Sushereba, C.[3], Ernst, K.[3], & Militello, L.[3]

5d. PROJECT NUMBER

Letter Agreement No. AS20-01-01501

5e. TASK NUMBER

03

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Applied Decision Science
1776 Mentor Ave., Suite 424, MB#118
Cincinnati, OH 45212

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

U.S. Army Aeromedical Research Laboratory
6901 Farrel Rd.
Fort Rucker, AL 36362

10. SPONSOR/MONITOR'S ACRONYM(S)

USAARL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

USAARL-TECH-CR--2022-22

12. DISTRIBUTION/AVAILABILITY STATEMENT

DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES

[1]Roth Cognitive Engineering; [2]Marimo Consulting; [3]Applied Decision Science

14. ABSTRACT

This report details the results of the Aviation Decision Making and Situation Awareness study. The objective of this study was to recommend measures and methods to evaluate future technologies that influence pilot decision making and situation awareness (SA) in the context of Future Vertical Lift (FVL). With regard to evaluation design, we offer two high-level recommendations. First, we recommend the use of scenario-based methods to test and evaluate technologies, with an emphasis on exploring a range of realistic scenarios, including cognitively challenging situations and 'edge cases.' Second, we recommend using multiple complementary measures to assess the impact of new technology on workload, SA, and other macrocognitive functions. Finally, we encourage the development of methods and best practices for evaluating integrated systems containing multiple technologies and person-technology interfaces anticipated for FVL cockpits so as to minimize potentially conflicting or inconsistent information.

15. SUBJECT TERMS

Decision making, situation awareness, SA, aviation, Future Vertical Lift, FVL

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	71	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

This page is intentionally blank.

Methods and Measures to Evaluate Technologies that Influence Aviator Decision Making and Situation Awareness

March 31, 2022

AUTHORS

Emilie Roth *Roth Cognitive Engineering*
Devorah Klein *Marimo Consulting*
Christen Sushereba, Katie Ernst, & Laura Militello *Applied Decision Science*

About Applied Decision Science

Applied Decision Science, LLC (ADS) is a research and development company focused on supporting human cognition in environments that are characterized by high stakes, high time-pressure and/or high complexity. Led by Senior Scientist and Chief Executive Officer, Laura Militello, ADS applies Naturalistic Decision-Making methods and models to support human performance in complex environments. Our emphasis is on studying how decision makers operate in the real world to articulate cognitive challenges and the skills and strategies that experienced operators employ to manage complexity. We rely on a suite of cognitive task analysis methods that drive the design of decision support tools, training programs and work-process redesign services to support and to improve decision-making. The company primarily provides its services in military and health care environments.

ADS leads a strong coalition of small businesses in this cognitive workload risk mitigation study. Roth Cognitive Engineering, founded by Dr. Emilie Roth, is a small company that conducts research and application in the areas of human factors and cognitive engineering. Marimo Consulting, established by Dr. Devorah Klein, focuses on design research and strategy for complex systems, grounded in cognitive psychology. For the past four years, ADS has led this coalition in several efforts related to Future Vertical Lift including the U.S. Army's Optimally Crewed Vehicle program that developed the Integrated Cognitive Analyses for Human-Machine Teaming (ICA-HMT) strategy and Cognitive Workload Risk Mitigation study.

www.applieddecisionscience.com

Acknowledgements

This material is based on work supported by the U.S. Army under agreement AS20-01501 Task 03. The authors would like to thank the U.S. Army Aeromedical Research Laboratory, and Combat Capabilities Development Command - Aviation and Missile Center - Technology Development Directorate for their sponsorship of this work. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Army. The authors would also like to thank Julie DiIulio and Eli Wagner, of Applied Decision Science, for their support in this study.

We are deeply indebted to Paul St. Onge, Grant Taylor, Larry Shattuck, Matt Shivers, Rob Moroney, Dave Durbin, and Jamison Hicks for generously sharing their time and expertise.

TABLE OF CONTENTS

Executive Summary

This report details the results of the Aviation Decision Making and Situation Awareness study. The objective of this study was to recommend measures and methods to evaluate future technologies that influence pilot decision making and situation awareness (SA) in the context of Future Vertical Lift (FVL).

For the purposes of understanding decision making in the FVL aviation environment, we use the following definition of decision making:

> *Decision making encompasses the cognitive activities involved in forming and refining a belief or course of action.*

In an interim report, we (1) reviewed current theoretical approaches to characterizing decision making and SA and (2) identified the implications of alternative theoretical perspectives for methods to operationally evaluate the impact of new technology on decision making and SA in the Army aviation domain. We created a synthesized model of decision making and SA by integrating core concepts from the decision-making models that are most relevant to FVL aviators. The interim report is available at the USAARL technical reports website.

The goal of this report is to support and guide people in the research, development, test, and evaluation disciplines as they create evaluation plans and select methods and measures to better assess the utility and efficacy of potential technologies for the FVL aviator. This report includes (1) a review of methods and measures for evaluating the impact of technology on decision making and SA, and (2) recommendations for evaluating technologies with respect to how they affect the five macrocognitive functions described in the synthesized model of decision making.

There were seven key contributions of this study.

Key Contribution 1: A review of current models of decision making and SA. In the first report, we reviewed and summarized models and theories from behavioral economics, cognitive psychology, human factors engineering, naturalistic decision making, and practitioner communities. These were examined for their relevance to FVL aviators.

Key Contribution 2: A synthesized model of decision making and SA. In the first report, we synthesized the core concepts of the models we reviewed, creating a combined model of aviator decision making and SA. The synthesized model of decision making is built on the key macrocognitive functions most relevant to FVL aviators (sensemaking, directing attention, managing workload, planning, and communicating/coordinating). The macrocognitive functions are foundational for both rapid, intuitive decision making and slower deliberative decision making. The model consists of two loops, Assessing and Acting, linked by sensemaking. Sensemaking refers to the process of integrating new data and existing knowledge to create an understanding of what is happening and to generate predictions of how the situation will evolve. The two loops represent dynamic processes that both inform sensemaking and result from it. The outputs include evolving plans, communications, and actions.

Key Contribution 3: Consolidation of cognitive requirements anticipated for FVL aviators. We analyzed findings from previous studies of the FVL domain to identify the anticipated key decisions of FVL aviators and developed a list of the key cognitive requirements for these aviators. We tied the cognitive requirements to the macrocognitive functions highlighted in the synthesized model of decision making and SA.

Key Contribution 4: Characterized list of emerging decision aids relevant to FVL aviators. We first identified emerging technologies designed to support Army aviators in aviation and navigation, communication, and advanced teaming. We then characterized how these aids support the five key macrocognitive functions underlying aviator decision making and SA.

Key Contribution 5: Lessons learned about how new technologies impact decision making. We summarized the lessons learned from the past 30 years of implementing new technologies. In particular, we highlighted positive and negative impacts of past technologies on operator decision making and SA. Understanding where person-technology systems have fallen short in the past provides important foundation for evaluating the effect of new technologies on decision making and SA.

Key Contribution 6: Experimental methods and measures for evaluating decision making and SA. We reviewed methods for studying decision making and SA in domains with similarities to Army aviation and identified measures for evaluating sensemaking, directing attention, managing workload, planning, communicating, and coordination. We provide recommended outcome-based performance measures, process measures, test participant assessments, and physiological measures for testing new FVL aiding technologies.

Key Contribution 7: Recommendations for evaluating the effects of new technologies. We provide a recommended process for designing evaluation studies to determine the effects of new aiding technologies on FVL aviator decision making and SA. Evaluation design should include articulating evaluation questions, designing the study, creating evaluation scenarios, and identifying measures. Each step of the recommended process should consider the known limits of both the technology and the human. Example evaluations of two hypothetical aiding technologies are also provided.

Recommendations: With regard to evaluation design, we offer two high-level recommendations. First, we recommend the use of scenario-based methods to test and evaluate technologies, with an emphasis on exploring a range of realistic scenarios, including cognitively challenging situations and ‘edge cases,’ Second, we recommend using multiple complementary measures to assess the impact of new technology on workload, SA, and other macrocognitive functions. With regard to next steps, we encourage USAARL to continue to codify, operationalize, and validate measures tailored for use in the FVL context. We recommend exercising the evaluation process outlined in this report to develop best practices for evaluating new technologies in terms of decision making and SA. We recommend creating opportunities to disseminate identified best practices through, for example, workshops and practitioner handbooks. Finally, we encourage the development of methods and best practices for evaluating integrated systems containing multiple technologies and person-technology interfaces anticipated for FVL cockpits so as to minimize potentially conflicting or inconsistent information.

1 Introduction

The U.S. Army is advancing new pilot decision-aiding technologies as it develops next generation Future Vertical Lift (FVL) rotorcraft and continues upgrading the modern Army Aviation fleet. The Army envisions using these rotorcraft in the joint all-domain operations environment, a battlespace that will place significant demands on the pilot. Furthermore, the demands of joint all-domain operations will differ significantly from the demands of counter-insurgency operations or conventional near-peer engagements of the past. More than ever before, the pilot will also serve as a mission manager, consuming data from distributed sensors, directing the employment of small unmanned aircraft called air launched effects (ALE), and communicating to achieve mission effects. Mission success will require expert decision making and high levels of situation awareness (SA).

Producing platforms that effectively support the pilot will require iterative development and testing to assess the capabilities, suitability, and usability of the system. Furthermore, these advanced systems will necessitate new methods and measures to assess how these technologies influence pilot decision making and SA. The U.S. Army Aeromedical Research Laboratory (USAARL) engaged with Applied Decision Science to review the scientific literature and recommend methods and measures to evaluate emerging technologies that will influence the decision making and SA of future pilots.

This study included two parts. In the first part of the study, the research team reviewed theories of decision making and SA. Concepts from behavioral economics, cognitive psychology, human factors engineering, naturalistic decision making, and practitioner communities were reviewed for their relevance to aviation decision making and SA. Drawing on this literature, the team articulated a *synthesized model of decision making* that emphasizes the macrocognitive functions that underly both rapid intuitive decision making and slower deliberative decision making. The *Aviation Decision Making and Situation Awareness Study: Decision Making Literature Review* report (Roth, Klein, & Ernst, 2021) summarized findings from that review. This report is available in the USAARL technical reports website.

In the second part of the study, the team undertook several streams of analysis that culminated in recommendations for methods and measures to evaluate technologies that influence aviator decision making and SA. The team analyzed existing cognitive task analyses and articulated the anticipated cognitive requirements of FVL aviators. Emerging decision aids relevant to FVL were identified and characterized with respect to the synthesized model of decision making. The team summarized lessons learned from previous efforts to develop decision-aiding technologies. Organized around the macrocognitive functions of the synthesized model of decision making, the team identified methods and measures including outcome performance measures, process measures, test participant assessment, and physiological measures. Finally, the team generated recommendations for evaluation methods and measures. To make these recommendations concrete, the team also generated two case studies describing how one might evaluate two different hypothetical aids: one to notionally support landing to exfiltrate ground forces, and a second to support commanding ALEs.

In this present report, we summarize the findings from the literature review and articulate findings from the review of methods and measures of decision making and SA.

- In Chapter 2 of this report, the theories of decision making and SA are summarized. Chapter two includes a definition of decision making and a refined version of the synthesized model of decision making drafted in the first report.
- Chapter 3 describes the types of decision making anticipated for future FVL pilots and highlights links between Army pilot decision making and the synthesized model of decision making.
- Chapter 4 identifies examples of emerging decision aids under development by the research community and the U.S. Army.
- In Chapter 5, we describe lessons learned from the implementation of technology to support decision making and SA.
- The output of the second part of our study, a review of methods and measures for assessing and evaluating decision making and SA, is presented in Chapter 6 of this report. Chapter 6 describes the methods and measures, organized around the five macrocognitive functions of the synthesized model.
- Chapter 7, Pulling It Together, includes our recommendations as well as an example to illustrate how one would evaluate a hypothetical landing aid designed to support pilots landing to exfiltrate ground forces.
- Chapter 8 offers conclusions and recommended next steps.

Our goal is that this report will be useful to researchers, developers, and evaluators as they strive to build, integrate, and validate systems that support and improve the decision making and SA of FVL aviators.

2 What Do We Mean by Decision Making? A Macrocognitive Perspective

Decision making has been widely studied by psychologists, philosophers, neuroscientists, mathematicians, economists, and others. As with many constructs, there are as many definitions as there are researchers. Early researchers understood decision making as the act of selecting between several alternative paths (Beach, 1993). Early decision-making research was done in labs, with college students making trivial decisions in artificial, but easy to control contexts. This approach produced carefully controlled studies of narrow phenomenon. Many early decision researchers also focused less on expertise and more on flaws in human judgement in performing this task (Kahneman & Tversky, 1979), giving birth to the heuristics and biases field.

More recently, naturalistic decision-making researchers have studied decision making from a different perspective, emphasizing the use of field methods to understand decision making in real-world contexts. The naturalistic decision-making movement studies decision making by experts in context, focused on high stakes decisions, often in dynamic conditions, with uncertainty, and under time pressure. There is also an emphasis on understanding what factors help people make great decisions, rather than how people are biased. It makes for more challenging study design, but with much more applicability to the FVL context. These researchers have found that experts don't always compare options and select the best; they may recognize a situation as similar to one they remember, evaluate if the solution to the prior problem would work in the present situation, and if so, use it. While this path might not always result in an optimal solution, experienced practitioners are generally able to quickly arrive at a solution that meets the demands of the situation and act in the time available. These researchers define decision making not as selecting between options, but determining a direction, which might not involve comparison at all. In this study, we have used this definition:

> *Decision making encompasses the cognitive activities involved in forming and refining a belief or course of action.*

REVIEW OF THEORETICAL MODELS

In the interim Aviation Decision Making and SA report (Roth, Klein & Ernst, 2021), we reviewed models of decision making in depth; we provide a brief summary here. There are three well established research-based models of decision making:

- ***Two System Model:*** This model focuses on two different modes of decision making: slow and deliberative or fast and intuitive (Kahneman, 2011). This model has contributed significantly

to our understanding of intuition and how decision aids can support better decision-making performance.

- ***Recognition-Primed Decision (RPD) model:*** According to RPD researchers, experts may not always make decisions in a rational way by comparing options with pros and cons (G. Klein, 1989). Instead, they intuitively recall prior situations they have encountered and if a solution to a similar situation worked before, they will use it again without considering other options. The naturalistic decision-making movement emerged from this way of thinking about decision making—that decision making should be studied in the real-world, looking at experts making real decisions.
- ***Situation Awareness (SA) model***: The study of SA came out of work on military aviation and is defined as, "the perception of the elements in the environment within a volume of time and space" (Level 1 SA), "the comprehension of their meaning" (Level 2 SA), and "the projection of their status in the near future" (Level 3 SA) (Endsley, 1995a). This model has been particularly valuable in its focus on the dynamic nature of decision making and highlighting the role of comprehension (also sometimes called sensemaking) in guiding decision making.

We also reviewed models from practitioner communities that focus on naturalistic contexts, including the:

- ***OODA Loop***: This model was developed by a military strategist, and moves from observing the world into orienting, deciding, and acting. The emphasis is on looping from observing and understanding (orienting) into action which informs where attention will be focused next (Boyd, 2018). Developed independently from academic research, it has many commonalities and overlaps. Particularly helpful is the focus on decision making as a continuous cycle.
- ***The Decision Ladder Model***: Originating in the process control industry, the Decision Ladder model described the processes used by expert operators in both routine and unfamiliar contexts. (Rasmussen, 1976; Rasmussen et al., 1994). This model has implications for the design of aids to support pilots and other operators in challenging situations.
- ***Macrocognition***: This more recent approach builds on prior models to describe individual and team cognitive functions in naturalistic settings, such as sensemaking, planning, coordinating, etc. (e.g., D. Klein et al., 2000; G. Klein, et al., 2003; Vicente et al., 2004; Shattuck & Miller, 2006). We found this approach a good foundation for developing a model that is relevant to the FVL program. While there are multiple models describing macrocognition, we have proposed a synthesized macrocognitive model of decision making that draws from all of the models discussed in the interim report.

We also discussed mathematical models of decision making, including Signal Detection Theory (Green & Swets, 1966) and the LENS model (Brunswik, 1955). Signal Detection Theory has utility in creating tools to aid in directing attention and detecting signals. The LENS model is a helpful tool to use in modeling how well a pilot has detected a set of cues and made an appropriate judgement.

Looking across all of these models, there are some common themes:

- The decision-making process may be intuitive, deliberative, or some combination of the two.
- Recognition-primed decision making, a more intuitive process, is a signature of expert performance, particularly in high stakes, time-pressured tasks.

- Other macrocognitive activities (including directing attention, sensemaking, planning and replanning) both influence decision making and are influenced by it, in a dynamic, iterative process.
- Perception is shaped both by prior expectation and detection of salient information.
- Sensemaking is an active process that is at the core of decision making. It is driven by previous knowledge but also shaped by the current realities. Part of sensemaking is creation of a story, or model, of what is happening (sometimes called a situation model or mental model).
- A situation model allows people to make predictions and plans about what might happen next, and also to revise plans as their models evolve.
- Strong collaboration and coordination with other team members depends on a shared model of the current situation and goals (sometimes called common ground, or shared situation awareness).

SYNTHESIZED MODEL OF DECISION MAKING

A primary output of our review of decision-making models, was a synthesized model that highlighted points of consensus on important macrocognitive functions that underpin decision making in complex dynamic environments such as FVL. In the upcoming chapters of this report, we focus on describing the relevance of this synthesized model for:

- Characterizing pilot decision making in FVL,
- Characterizing emergent technologies intended to support pilot decision making in FVL, and
- Identifying measures and methods that can be used to evaluate the impact of new technologies of pilot decision making in FVL.

This model was synthesized from the broad and diverse research on decision making, and largely influenced by a macrocognitive perspective. The aim is to create a model that reflects the strong research already done, and also takes into account the challenges of the FVL context in which it will be used. The synthesized model of macrocognitive functions underlying complex decision making is presented in Figure 1. This figure is a refinement of the version that appeared in the interim report (Roth, Klein & Ernst, 2021). While the core concepts underlying the model remains the same, the figure depicting the model has been revised in response to stakeholder feedback.

We summarize the model here to ground the rest of this report; we anticipate that this model will be used as a tool to guide evaluation of future technologies.

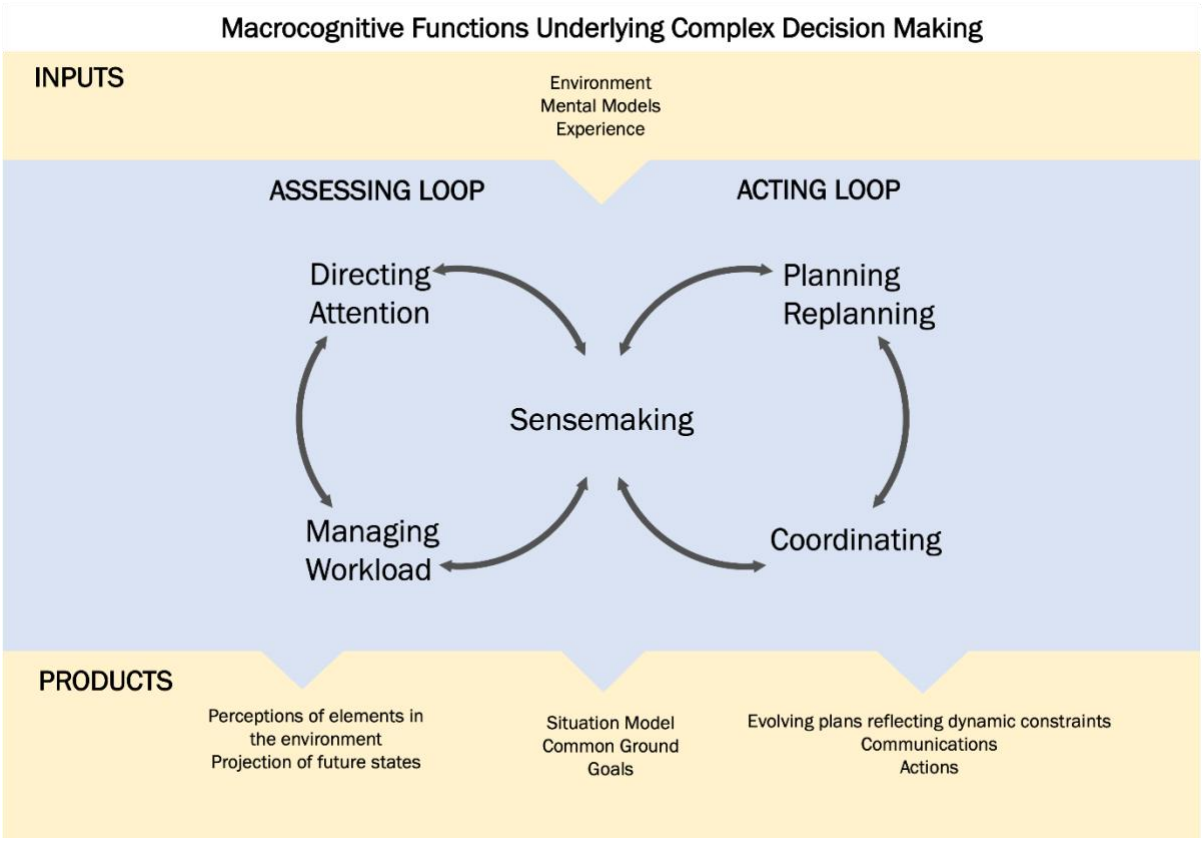

Figure 1 A synthesized model of decision making

This synthesized model of decision making is built on the key macrocognitive functions most relevant to the FVL program. It consists of two primary loops, Assessing and Acting, linked by sensemaking. The phrase "loops" is intended to highlight the iterative and dynamic nature of these processes. We begin by focusing on sensemaking, due to its importance as the linchpin connecting them.

Sensemaking

Sensemaking links the Assessing and Acting loops; sensemaking is the process of integrating new data and existing knowledge to create an understanding of what is happening to generate predictions and plans for the future. Sensemaking may be deliberative and slow; it may be intuitive and recognition primed, but it is always shaped by stored knowledge and mental models as well as the current environment. Sensemaking turns assessment into action and turns actions into a search for more information. This centrality is particularly true for pilots in the FVL context and so to support better decision making for pilots, it is critical to support sensemaking and the related functions (for more discussion of sensemaking, see Roth, Klein, & Ernst, 2021).

Assessing Loop

Often, the starting point for making a decision comes from new information that is gathered and deemed important. This is the focus of the Assessing Loop. In addition to Sensemaking, Assessing consists of:

- ***Directing Attention***: This macrocognitive function includes noticing things happening in the world, as well as searching for cues and information. Attention can be both a bottom-up process (seeing something interesting or different) and top-down process (actively searching for information).
- ***Managing Workload***: Another key aspect of macrocognition is how pilots manage their workload, what they focus on, and what they intentionally ignore (at least for a while). Experts can fluidly triage what is important, what can be set aside for the moment, and what can be shared to a teammate as they move between tasks.

Acting Loop

The Acting Loop is where sensemaking guides action and decision. It is also where the acts of planning and coordinating can feed back into sensemaking and the Assessing Loop if more information is needed. Planning and making decisions are rarely a once and for all activity; rather plans get dynamically revisited and revised as a situation unfolds.

- ***Planning/Replanning***: Planning covers both simple intentions to act (e.g., determining to fly higher to avoid an obstacle) as well as formulating complex plans to achieve goals (e.g., generating an elaborate course of action for accomplishing a mission). Determining a plan and then changing it as more information is learned is a central part of what pilots do—they are constantly evolving their decisions as their understanding of the situation changes and priorities shift. Part of planning is also how pilots manage uncertainty, sometimes making a temporary decision and seeking more information.
- ***Coordinating***: For pilots, coordinating with others is another significant part of their job—whether it is others in the helicopter, team members on the ground, and even the technology itself, communicating and coordinating is essential to developing shared sensemaking that leads to common ground. By common ground, we mean a shared understanding of the situation, as well as goals and priorities (G. Klein et al., 2005).

Inputs and Outputs

It is also important to discuss the inputs to the process and the outputs. Key inputs to the decision-making process include the environment itself, and the mental models and experiences of the pilots and other stakeholders. These shape what information is observed, how it is understood, and the goals and priorities that in turn influence planning and replanning.

Outputs of the process include:

- Perception of elements in the environment (Level 1 SA).
- An understanding of the current situation (a situation model) that is derived from the sensemaking activities. This situation model can also be thought of as Level 2 SA – comprehension of the situation.

- Projection into the future (Level 3 SA).
- Identification and or reprioritization of goals that may lead to further planning activity, and/or seeking additional information through directing attention.
- Evolving plans that reflect dynamic constraints and uncertainties.
- Actions.

When operating as part of multi-person teams, outputs would also include:

- Common ground – a shared understanding of the situation and goals and priorities across multi-person teams.
- Communication and coordinated action.

In the following chapters we leverage this synthesized model to discuss strategies for supporting pilots in the future: how they make decisions, how emergent technologies can support them, and most importantly, how to assess the impact of emergent technologies on these macrocognitive functions.

3 Pilot Decision Making in FVL

Decision making in Army Aviation is complex, given the use of high-performance rotorcraft flying in close proximity to terrain and vertical obstructions, often in degraded visual environments, under rapidly changing conditions, and in complex battlespaces. As a principle, the Army values good decision making. The Army invests many flight hours and trains aviators on the military decision-making process and the rapid decision-making synchronization (Headquarters, Department of the Army, 2019) process to build the reflexive and adaptive skills needed for sound decision making on the battlefield. As we look at evaluating technologies that will influence pilot decision making and SA, the questions therefore are:

- *What does decision making in Army Aviation look like?*
- *Does the synthesized model accurately represent aviator decision making?*

WHAT WILL DECISION MAKING LOOK LIKE FOR FVL AVIATORS?

Over the past few years, our research team has conducted multiple cognitive task analyses examining the key decisions of Army aviators in a variety of missions (Militello et al., 2018; Militello et al., 2019a; Militello et al., 2019b; Ernst et al., 2021). We used cognitive task analysis to identify the cognitive demands and skills relevant to particular tasks, as well as links to critical cues, expectancies, and goals necessary for good decision making in specific contexts. This work yielded a corpus of cases that describe real-word events, the key decisions aviators made in those situations, and a set of key cognitive requirements relevant to FVL.

In addition, in 2020 our research team conducted a study of the Army's efforts to mitigate the risk of cognitive overload in FVL aviators (Ernst et al., 2020). The study identified seven key performance attributes of FVL that would influence workload and impose new or increased cognitive demands on Future Armed Reconnaissance Aircraft (FARA) and Future Long Range Assault Aircraft (FLRAA) aircrew. FVL aircrew can expect to fly aircraft with increased agility in degraded visual environments, send and receive high volumes of SA data, use sensor data from fused and emerging sensor capabilities, and operate in a complex threat environment at fast-paced operations tempo.

At a high level, these cognitive task analyses revealed that aviators are and will be constantly making decisions, large and small, from the moment they receive a mission tasking to when they land the helicopter after the mission. The incident accounts in the corpus of cases include blink-of-the-eye events, such as when a pilot notices a ball of light, instinctively recognizes a rocket propelled grenade, and turns his aircraft away from the threat, as well as more deliberative decision making where a pilot gathered information about a situation over tens of minutes in coordination with other aircraft before taking action. Looking across the four studies examining a variety of

missions and tasks, Table 1 includes key cognitive requirements needed to support decision making relevant to FVL operations.

Table 1 Key cognitive requirements anticipated for FVL and associated macrocognitive functions

Cognitive Requirement[1]	Sensemaking	Directing attention	Managing workload	Planning	Coordinating	Source[2]
Developing & maintaining a 3D understanding of the battlespace	✓	✓				c
Developing expectations pre-flight	✓	✓	✓	✓	✓	c
Correlating representations & real world	✓	✓				c
Perspective taking	✓				✓	c
Understanding mission goals to answer priority intel requirements	✓			✓	✓	r
Send & receive SA data to maintain common operating picture	✓	✓	✓	✓	✓	a
Managing sensors & sensor data	✓	✓	✓	✓	✓	r
Managing automation & supervising autonomy	✓	✓	✓		✓	cw
Recognize opportunities	✓	✓				w
Detecting & locating objects in real time	✓	✓		✓		ca
Tracking targets	✓	✓			✓	ra
Reacting to objects in real time	✓	✓		✓		c
React to a threat in real time	✓	✓				c
Battle damage assessment	✓				✓	a
Dynamic replanning	✓	✓	✓	✓	✓	ra
Minimize detection & maximize survivability	✓			✓		w
Knowing one's aircraft capabilities for current conditions	✓	✓				c
Spatial awareness & positioning	✓	✓				ca
Maneuvering for complex pilotage tasks	✓	✓		✓		w
Flying in reduced cueing environment	✓	✓	✓			c
Maintain & coordinate mission information and situation understanding across planning team and the aircrew	✓				✓	w
Disseminating timely & appropriate information	✓			✓	✓	cra
Manage interpersonal dynamics	✓				✓	cra
Dealing with off-nominal conditions	✓	✓	✓			ca
Adapt to different roles and workflow	✓	✓	✓	✓	✓	w

Table note: [1] The listed cognitive requirements are organized in groups of similar or associated cognitive requirements. Groups are separated by blank rows. [2] Source indicates the cognitive task analysis that generated this cognitive requirement. Key: a – Militello et al. (2019a); r – Militello et al. (2019b); w – Ernst et al. (2020); and c – Ernst et al. (2021).

DOES THE SYNTHESIZED MODEL ACCURATELY REPRESENT AVIATOR DECISION MAKING?

The five main functions of the synthesized model, sensemaking, directing attention, managing workload, planning, and coordinating, are high level components of the cognitive requirements identified in Table 1. Each of the cognitive requirements was mapped to one or more macrocognitive functions in Table 1. Looking at these cognitive requirements and FARA and FLRAA operational concepts, we posit that these cognitive requirements capture the cognitive demands facing Army aviators in joint all-domain operations. Figure 2 depicts, in general, how the macrocognitive functions support aviator decision making in the context of a mission.

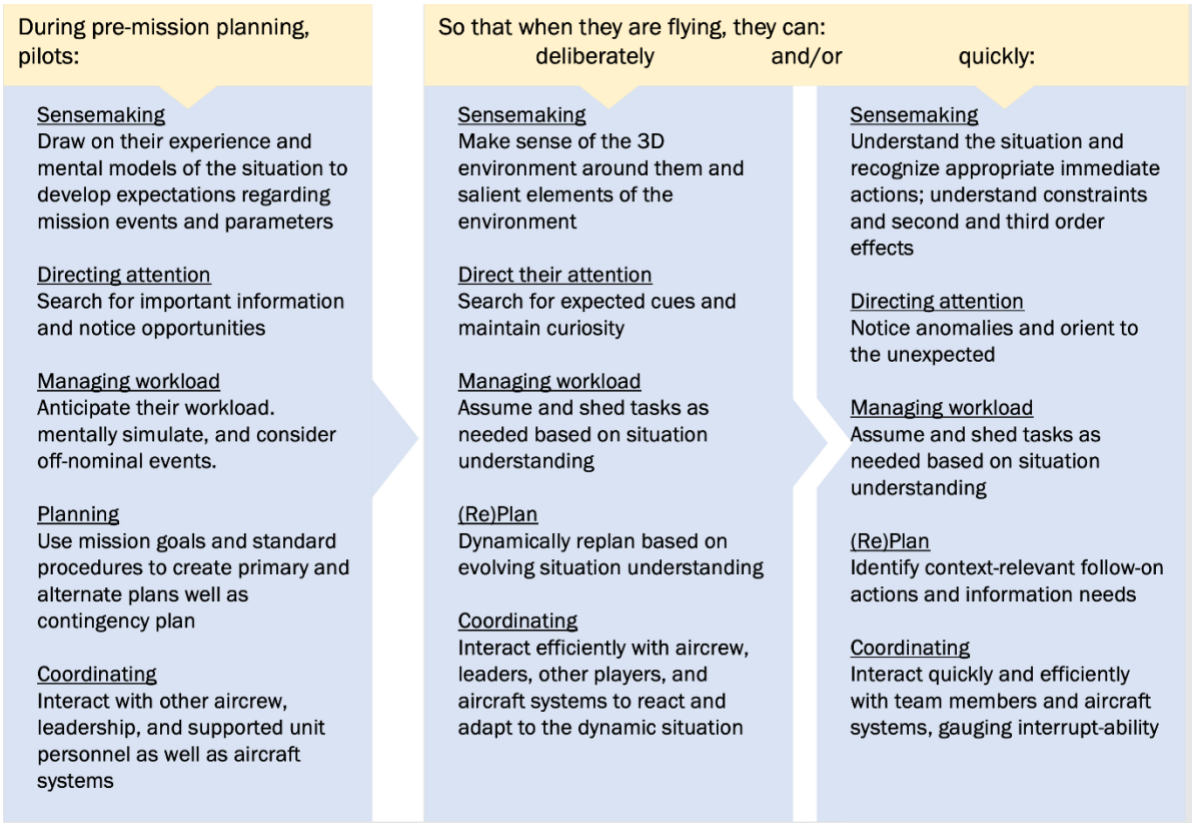

Figure 2 Macrocognitive functions in the context of a mission

First and foremost, aviators will need to ***make sense*** of the unfolding mission situation, ***directing their attention*** between various inputs such as information from ALEs and fused data streams. This sensemaking activity will require understanding the commander's intent, the enemy force, and the needs of the supported unit. Aviators will also ***manage their workload***, supervising automated and autonomous on- and off-board systems, offloading the right tasks to

automation/autonomy to maintain SA and an appropriate workload. Aviators will be dynamically ***planning*** and replanning, updating the plan to accommodate the changing battlefield environment. Finally, aviators will be ***coordinating*** with other humans (aviators, leadership, supported units, joint and allied forces) and automated/autonomous systems to exchange information, synchronize actions, and communicate commands.

An incident recounted in the Complexity in Information Systems program (Ernst et al., 2021) illustrates the types of decisions aviators face today and will likely face in the future. An experienced Apache pilot described a mission where he served as Attack Weapons Team Leader (AWT) for a two-ship of AH-64D Apaches supporting a ground force. In this operation, the ground force's plan included infiltrating from the edge of a mountainous village and moving through the village guided by the developing intelligence situation. As the four Apache pilots planned the mission, they examined the maps, imagery, and gridded reference products of the village to identify potential areas of concern and courses of action. In planning, they identified an overhang/cave overlooking the village that could provide a strong defensive position to an adversary and a threat to the ground force. The pilots planned their orbit to allow them to overwatch the entire village at a safe distance from the terrain and vegetation. Once airborne with the ground forces in the village, the AWT leader noticed that the ground forces were moving up toward the overhang and recognized that his planned orbit (that fully utilized the onboard hold modes) would not allow his aircraft to maintain sensors on the target. However, he determined that a narrow band of airspace and a complex manually flown geometry would allow him to stay away from vertical obstructions and in view of the activity beneath the overhang at all times. The AWT leader coordinated with the other Apache pilot to manually fly a series of turns that increased the workload of the flying task but allowed him to successfully support the ground force.

This account provides a glimpse at the decision making facing FVL aircrew. FVL pilots will be supporting combined arms maneuvers using advanced flight control and mission systems to sense the environment to make tactical decisions. In this account, during pre-mission planning, the crews ***made sense*** of the supported unit's plan and the terrain, and drew on mental models of their own aircraft's performance, enemy and friendly tactics, and capabilities. They ***directed their attention*** between various information sources searching for features while noticing items important to themselves as aviators and the supported unit. Using this information, the crews ***planned*** out the mission and ***coordinated*** with each other, the lift crews, and the ground force. Once in the air with ground forces at the objective, as he was ***directing his attention*** between his own system displays and the co-pilot-gunner systems, the AWT leader recognized the ground forces heading toward the overhang. His ***sensemaking*** of the emerging situation allowed him to rapidly ***replan*** his own flight path. In this plan, he ***managed the workload*** across his crew, planning an orbit which was easier for the copilot/gunner to maintain eyes on the target, even as it induced higher workload for himself. Finally, he ***coordinated*** with his co-pilot/gunner and sister ship pilot using voice and tactical data link communication methods.

The macrocognition-based model of pilot decision-making represents the interconnected and iterative nature of macrocognitive functions associated with decision making and SA. Moreover, the model, oriented around sensemaking, reflects a key challenge that FVL aviators will face, making sense of the environment. Maintaining SA and making good timely decisions will require effective support as pilots consume an ever-increasing amount of data and information while managing effects and flying their own aircraft in a complex battlespace.

4 Emerging Technologies to Support Pilot Decision Making

There is an accelerating effort in government, academia, and industry to develop aiding technologies to support increased performance for advanced rotorcraft platforms and aircrews. Decision-aiding systems include front-end human machine interfaces and back-end technologies. While some development efforts are relatively mature and integrated (i.e., at higher technology readiness level), other important aiding systems are still at a basic research or lower technology readiness level. Technologies in development include those designed to aid pilots in the five primary mission tasks (e.g., aviate, navigate, communicate, managing sub-systems status, and weapons engagement) as well as new tasks related to mission systems such as advanced teaming.

We reviewed a selection of Science, Technology, Research, Engineering, Test, and Evaluation technologies under development. In this chapter we will describe those development efforts and describe how these technologies may support FVL aircrew decision making and SA. These technologies are organized around the following categories: aviation and navigation aids, communication aids, and advanced teaming aids.

AVIATION AND NAVIGATION AIDS

The recently completed Army Degraded Visual Environment-Mitigation (DVE-M) program facilitated integrated development of two systems that support pilots in aviating and navigating in degraded visual environments. The Mission Adaptive Autonomy (MAA) and Integrated Cueing Environment (ICE) together support pilots by providing guidance and control for manual, aided, and autonomous flight, as well as pilot vehicle interfaces for obstacle detection and navigation (Szoboszlay et al., 2021; Miller et al., 2021; Whalley et al., 2014; Takahashi et al., 2021). MAA and ICE supports pilot decision making and SA in a number of ways.

- ***Managing workload:*** The systems allow the pilot to manage their workload by relying more or less on the automated systems. In flight test trials using a full-authority Black Hawk, the system flew autonomously while giving the pilot the ability to bias the flight path (Takahashi et al., 2016). In later flight test trials on a partial authority Black Hawk, pilots were able to fly autonomously, aided by the path planning symbology, or standard coupled modes (Takahashi et al., 2021). Reducing workload was a priority for the DVE-M program, and the technologies developed allow the pilot to delegate more or less authority to the autonomous systems.

- ***Directing attention:*** The ICE symbology and multi-modal cuing are the product of decades-long research on cuing in degraded visual environments. These systems alert the pilot to hazards in the nearby airspace by presenting artificially rendered obstacles in color against the grey scale background terrain image (Szoboszlay et al., 2021). Conformal symbology alerts the pilots to landing pad locations (Figure 3). In trials using bumper-radar systems, spatial-auditory and tactile cueing alerted and warned pilots of vertical obstructions (Miller et al., 2021).

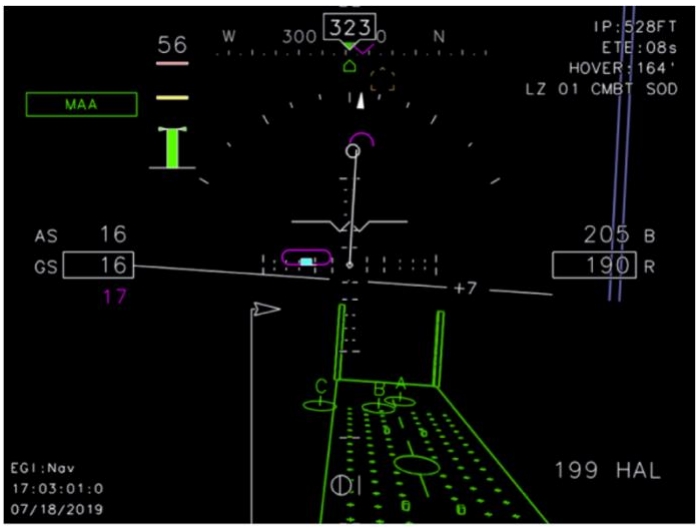

Figure 3 from Szoboszlay et al. (2021) depicting artificial landing pad symbology with options A, B, and C as identified by the safe landing area determination algorithm.

- ***Planning:*** MAA's flight path planner, Risk Minimizing Obstacle Field Navigation, uses a constraints-based algorithm that builds and updates a flight path based on digital terrain elevation data, on-board LADAR sensors, and other inputs (e.g., threats, restricted airspace) (Takahashi et al., 2021). This planning feature supports the pilot in pre-mission planning and in dynamic replanning events.
- ***Collaborating:*** The MAA Safe Landing Area Determination (SLAD) system supports communication between the automation and pilot. The autonomy scans the pre-designated landing zone and offers the pilot several options for landing based on pre-defined constraints. The pilot can then select an option, re-run the options, or manually land the aircraft supported by the symbology (Takahashi et al., 2018).
- ***Sensemaking:*** The conformal elements of ICE support pilot sensemaking. Rendering vertical obstructions as a layer over the background imagery supports pilot understanding of the environment surrounding the aircraft. The safety corridor display prototype (Szoboszlay et al. 2021) offered an alternative way to support the pilot in making sense of the dynamically updating flight path planner amid the terrain and airspace constraints.

Work conducted by researchers at Delft University tested two HMI displays supporting pilots in navigating over obstacles (Friesen, Borst, Pavel, Stroosma, Masarati & Mulder, 2021). This technology was designed to aid pilots' decision making in noticing obstacles and deciding when and how to navigate over vertical obstructions while accounting for aircraft parameters, distance from the obstruction, and height of the obstruction. Findings from this experiment demonstrated that pilots preferred a more directive HMI in nominal conditions and a less directive HMI in off-nominal conditions.

- ***Sensemaking:*** The directive display supported pilot sensemaking by displaying both the obstacle, and the need to ascend based on the obstacle's distance and height, and aircraft parameters. The less directive display depicted a conformal line indicating the minimum safety line *and* the maximum ascent possible at that heading, which was more useful to pilots in off-nominal conditions.

- ***Managing workload:*** This display concept highlights the utility of adapting the display elements to support pilot workload when in navigating complex obstacles.
- ***Directing attention:*** The directive display allowed the pilots to divert their attention to other elements of the display yet still be cued to the need to ascend.
- ***Planning:*** The directive display supported the pilot in offloading the ascent planning task to the automation.

COMMUNICATION AIDS

Decision aids under development will also support communication between humans and between humans and automated/autonomous systems.

An example of this is the dynamic infographic concept being developed as part of the Joint Health Services' decision-support platform, called Medical Common Operating Picture (MedCOP) program. Work currently being conducted by the Data and Analysis Center focuses on supporting pilot coordination with external entities, such as medics at a casualty collection point or aviation maintainers at a forward arming and refueling point (FARP; Hartnett & Hicks, 2021; D. Durbin & J. Hicks, Data and Analysis Center, interview with authors, January 2022). Today, information is passed as a string of text or verbally in a radio call, whereas dynamic infographics provide the content visually as a standardized dynamically updating graphic on the aviator's helmet mounted display. The infographic is connected with a conformal position indicator.

- ***Sensemaking:*** First, displaying tactically relevant dynamically updated data as an infographic supports aviators in making sense of the content. Rather than parsing spatial information passed via the radio or text message, the pilot sees a visualization of the underlying situation. For example, the infographic displays refueling positions as red or green parking spots as they are oriented at the FARP, allowing the aviator to rapidly make sense of the FARP's status. Second, by associating the dynamic infographic with a conformal element in the pilot's helmet mounted display supports the pilots understanding of elements in the 3D world.
- ***Managing workload:*** Passing information via multiple channels supports the pilot in managing their own workload. The number of voice communications channels in Army Aviation platforms already exceeds a person's ability to reasonably consume all the information. Sharing this information via visual-spatial and visual- channels allows the pilot to distribute their load across multiple senses.

Another example of an aid supporting communication between humans and automation is the Operator State Monitoring program. This program is developing a physiological model of aviator workload. Using this model of workload, the US Army Aeromedical Research Laboratory aims to support adaptive automation, that is, automation that autonomously selects when and how to adapt based on its perception of the pilot's state.

- ***Collaborating:*** This decision support aid will autonomously adapt the type of automation support based on the sensed physiological state of the pilot. When under high workload, the pilot may not be able to articulate their need for support or the type of support they require. At the other end of the spectrum, a pilot who is underloaded may not be able to articulate their level of boredom or disengagement. The use of physiologic measures offers a novel

mechanism for the automation/ autonomy to communicate with the pilot and provide a supportive amount of aiding.

ADVANCED TEAMING AIDS

Advanced teaming between manned platforms and many unmanned platforms is a new paradigm for Army Aviation. Recent programs such as the Army's Synergistic Unmanned Manned Intelligent Teaming (SUMIT) explored the impact of pilot aids that facilitated control of multiple unmanned aircraft (Alicia et al., 2020). Army Aviation development's primary effort in this area is Advanced Teaming (A-Team) which is developing software enabled, decentralized, and distributed command of multiple unmanned aircraft systems (UAS) by a single rotorcraft pilot (Aviation Development Directorate - Eustis, 2018).

A related effort developed jointly by the research laboratories of the Air Force, Army, and Navy is the IMPACT (Intelligent Multi-UxV planner with Adaptive Collaborative/control Technologies) program, designed to support a single ground-based operator manage 12 or more air, ground, or sea surface-based unmanned vehicles (Draper et al., 2018). The IMPACT integrated control station includes a cooperative control algorithm, intelligent agents that support allocating assets and identifying opportunities for action, autonomics to monitor the plan while executing, and human machine interfaces. A key feature of IMPACT is its use of "plays" for communicating tasking intent and priorities between humans and automated

- ***Managing workload:*** IMPACT reduces the workload associated with tasking and monitoring heterogenous air, ground, or surface unmanned vehicles (UxVs) by both delegating some or all of the task work to autonomy while facilitating human input at multiple levels (Draper et al., 2018). Experiments with IMPACT evaluated performance-based adaptive automation that automatically escalated the amount of support when an operator's performance declined (Calhoun, Bartik, et al., 2021).
- ***Directing attention:*** The autonomic monitoring system oversees the mission execution of UxVs and alerts the operator to off-nominal or context specific situations that may require human intervention (Draper et al., 2018).
- ***Planning:*** The use of "plays" and a "playbook" approach to tasking facilitates collaboration between humans and autonomy, allowing the human to pre-designate plays or parameters of interest, and adapt plays on the fly (Calhoun, Ruff, et al., 2021).
- ***Collaboration:*** Once a play has been called, either by the human or the autonomy, a standardized visualization displays the parameters on which the play was based (Calhoun, Ruff, et al., 2021). This visualization supports the operators understanding of the autonomy, as well as a platform for the human operator to influence the autonomy to change the play.
- ***Sensemaking:*** IMPACT's HMI includes a Solution Rationale Window that displays multiple Courses of Action (COA) that have been proposed, each optimized for a different parameter (Calhoun, Ruff, et al., 2021). This interface supports the operator in making sense of the autonomy's recommended COAs. In addition, the operator can use the interface to explore the solution space and iteratively refine presented options to arrive at a more suitable COA.

Researchers at Bundeswehr University Munich and HAT.tec GmbH are developing autonomous agents that support rotorcraft pilots conducting advanced teaming with unmanned assets (Brand &

Schulte, 2021). Their integrated suite of tools allows a helicopter crew (pilot in command and pilot flying) to fly a transport helicopter route under threat while managing a fleet of UAS conducting route reconnaissance. An intelligent cognitive agent associate system predicts pilot mental workload using an eight-dimensional demand vector that predicts upcoming pilot resource demands (using multiple resource theory) based on a task decomposition. The associate system uses the mental workload prediction and environmental context to adaptively assist the pilot with missed tasks or in critical states that require intervention.

- ***Directing attention:*** The associate system aids the pilot by identifying missed work and adaptively offering support based on the mission/situation context. At the lowest amount of intervention, the associate guides the pilot's attention to the problem, displaying the relevant information on the pilot's display and can display a shortcut button to execute the required actions.
- ***Managing workload:*** At a higher level of intervention or in situations with time-urgency, the associate system can opt to take over the missed task(s) or critical action(s).
- ***Collaborating:*** When the associate system adopts a task, it displays the task in a list of adopted tasks on the Multi-Function Display (MFD) and makes a synthesized verbal announcement as a human crewmember would do in the circumstance.

Looking across this set of aids, we can make a few observations. First, while these aids were designed to support the pilot in different tasks, we were able to examine all of the aids using the synthesized model of decision making. Second, the aids did not necessarily need to support all five of the macrocognitive functions to support decision-making and SA. For example, aids that support pilots in managing their workload may indirectly free attentional resources that allow the pilots to improve performance of the other four macrocognitive functions. Finally, the synthesized model served as a useful framework to assess how similar aids supported pilot decision making and SA in different ways.

5 Impact of Technology on Decision Making: Some Lessons Learned

The previous chapter described a variety of technologies that are emerging to improve FVL pilot SA and decision making. The objective is to augment human capabilities to produce better, faster, and more resilient performance in the face of fast paced challenging conditions. Lessons learned over the last 30 years has revealed that, unless carefully designed to take into account the needs for effective human performance, new technology will not necessarily improve performance and may even lead to poorer outcomes (National Academies of Sciences, Engineering and Medicine, 2021). Here we summarize some of these lessons learned, highlighting the need for carefully designed person-in-the-loop evaluation studies that incorporate appropriate measures to ensure that new technologies realize their intended benefits. The ultimate aim is to produce technologies that augment human capabilities and raise performance of the *joint person-technology system* beyond that of either entity working alone (Woods & Hollnagel, 2006).

A consensus report issued by the National Academies of Sciences, Engineering and Medicine (2021) highlights the range of known human cognition and performance problems that can arise from poorly designed automation. These include:

- ***Poor SA and out-of-the-loop performance degradation***: The introduction of automation can result in people becoming 'out-of-the-loop' meaning they are less likely to be aware of critical information and be able to take manual control when needed. The out-of-the-loop problem can result in severe performance consequences especially when confronting 'black swan' situations that are novel and unanticipated (Sebok & Wickens, 2017).
- ***Automation surprise***: A common problem with new technologies is that people fail to understand how the automation works leading to inaccurate expectations and inappropriate actions that can have catastrophic effects. These types of automation surprises are well-documented in advanced cockpits (Endsley, 2019; Rankin, et al., 2016; Sarter & Woods, 1995).
- ***Ironies of automation***: When technology is working well, people grow to rely on it and may lose the cognitive and manual skills to recognize when they need to take over and perform effectively. When situations beyond the capabilities of the technology arise, high workload spikes will occur, overstretching human performance (Bainbridge, 1983). At the limit a technology may completely shut down, with little warning or communication about why and what it was doing. The human is thrust into a high-stakes, stressful situation with a partner that has gone silent (Norman, 1990).
- ***Automation bias***: The introduction of technology aids can improve performance when the recommended solution is correct, but when it is wrong, it can lead to worse performance than the person working on their own (Smith et al., 1997). For example, Metzger and Parasuraman (2005) found that air traffic controllers performed better without an aid than with an imperfect conflict-detection system. Similar results have been shown repeatedly, including recently in a

study by Friesen, Borst, Pavel, Masarati, and Mulder (2021) who compared alternative advisory system displays for safe path planning in a helicopter flight route planning application. They report that when the advisory system generated a specific flight path, pilots tended to follow it, even when there were better trajectories available that would save fuel and time. In contrast, when the advisory system used constraint-based displays that made it possible to see multiple path options, pilots were more likely to select a more optimal route.

A large body of research has led to a general consensus that for human technology systems to be more effective than either working alone the person must be able to (1) understand and predict the behavior of the technology; (2) develop appropriate trust in the technology – knowing when it is likely to perform well and should be trusted, and when it is likely to be outside its bounds of competence; and (3) exert control over the system so as to redirect it toward a more productive path or takeover manually (Boardman & Butcher, 2019; National Academies of Sciences, Engineering and Medicine, 2021).

Researchers have been exploring ways to make technology more *transparent* (also sometimes called *observable*) so that its behavior is more understandable and predictable (e.g., Mercado, et al. 2016; Roth, et al., 2017). Increased transparency has been shown to improve SA, improve calibration of trust, and improve performance (Stowers et al., 2017).

The SUMIT program highlighted the importance of transparency on user acceptance and performance. The SUMIT program defined transparency as interface features that enable operators to understand the technologies intent, performance, future plans, and reasoning (Alicia et al., 2020). They report that participant feedback reinforced the importance of automation transparency. Examples of transparency in the SUMIT program include "depicting the planned flight path for all assets, presenting asset task assignments and status, and informing the user of any constraints that might prohibit task assignment, such as an unarmed asset unable to engage a target or a fast-moving asset unable to make a tight turn to follow waypoints spaced too close together." (Alicia et al., 2020, p. 20). They noted that "the inverse of automation transparency, which is automation opacity, resulted in significant frustration for the participants and often led to them to cancelling automated tasks and attempting to force the task via manual control without understanding the system reasoning that prevented task completion" (Alicia et al., 2020, p. 20).

There has also been considerable research in making technology more *directable.* Directability refers to the ability to direct and redirect the technology as needs or contexts shift (Christoffersen & Woods, 2002; Klein et al., 2004; Roth et al., 1987). There have been multiple efforts to develop more directable systems by enabling users to communicate goals, priorities, and constraints as well as directly influence its solution path (e.g., Johnson et al., 2018; Roth et al., 2017). At the minimum is the ability to override the technology and manually take over. Researchers from the SUMIT program report that the most discussed capability that arose during the evaluation was the need for participants to take immediate direct manual control for critical tasks such as sensor management and target engagement (Alicia et al., 2020).

A primary motivation for increasing observability and directability is to make the joint person-technology system more resilient in the face of unforeseen conditions (National Academies, 2021; Neville et al., 2021; Roth et al., 2019; Woods, 2015). Resilience refers to the ability to adapt to changing conditions, particularly conditions that are unanticipated and beyond the capabilities of

the aiding technology. Hoffman and Hancock define resilience as the 'capacity to change as a result of circumstances that push the system beyond the boundaries of its competence envelope. The system may have to amend some or even all of its goals, procedures, resources, roles or responsibilities' (Hoffman & Hancock, p. 565-566). Conditions that require resilience are likely to arise in FVL where unanticipated threats, rapidly changing priorities, and/or sudden system malfunctions or loss of resources (e.g., loss of ALEs, loss of communication) will require dynamic reassessment and replanning, and even novel response to meet mission objectives. In those conditions, it will be important for the joint person-technology system to be able to respond adaptively in order to successfully meet mission objectives. As a result, when evaluating new technology, it is important to understand how it impacts the ability of pilots to respond adaptively under unanticipated conditions, particularly conditions where the automated aid fails or generates suboptimal solutions.

Additional guidance on technology design features that support human performance can be found in Wiggins and Cox (2010) who cover a variety of considerations that range from system's ability to support users when considering options to the system's ability to adapt in a situation with changing priorities and goals.

The main purpose of this brief review of lessons learned from prior experience with automation and intelligent systems is to highlight that technology can have negative as well as positive effects on human performance. As a consequence, it is critical to conduct person-in-the-loop evaluations to understand the impacts of the technology on the macrocognitive functions that underlie decision making so as to identify and mitigate any potential problems early in the design cycle. The next chapter presents methods and measures that can be used in evaluating the impact of new technologies on SA and decision making.

6 Evaluating the Impact of Technology on Pilot Decision Making: Methods & Measures

New technologies are often introduced with the promise of improving performance, but the anticipated benefits are not always realized. It is consequently important to conduct person-in-the-loop evaluations of new technologies to establish that the anticipated benefits are realized and that there are no unanticipated negative impacts on performance. Person-in-the loop evaluations of new technologies should be conducted throughout the system life cycle starting at initial concept development, through system design and integration, and all the way up to and through system fielding (National Research Council, 2007; National Academies, 2021). Early evaluations may be conducted using rapid prototypes in laboratory settings, later evaluations may be conducted with integrated systems using high fidelity evaluations, and final evaluations should be conducted on the actual fielded system to assess how well it performs when confronted with the demands of the actual context of use.

When evaluating new technologies intended to support pilot decision making, it is important to assess how the technology impacts the different macrocognitive functions that underly decision making. Does it improve SA? Does it support sensemaking? What effect does it have on the pilot's ability to manage and direct attention to what is currently most important? How does it affect workload? The goal is to establish the envisioned benefits of the technology as well as check for potential negative impacts.

OVERVIEW

In this chapter we summarize measures and methods that can be used to evaluate the impact of new technology on the different macrocognitive functions that underlie decision making as well as on overall performance of the individual or team as supported by the technology (referred to as the joint person-technology system). The measures and methods summarized are based on a review of the literature as well as interviews with researchers supporting Army R&D Programs such as the A-Team program and the Holistic Situational Awareness and Decision making (HSA-DM) program.

There are a range of measures that can be used to evaluate macrocognitive performance. These include:

***Outcome Performance Measures*:** These measure the actual performance on the task of interest (e.g., did they detect the target? Did they select the right course of action?). Outcome measures typically include the quality of the response (e.g., error rate) as well as the time to come up with the response (response time).

Process Measures: These measure the cognitive and team processes that result in the outcome performance. Measures of macrocognitive functions generally fall in this category and include measures such as did they focus on the right information? Did they correctly understand the situation? Did they communicate effectively? Did they know to over-ride the recommendation provided by the technology in situations when it was wrong?

Test Participant Assessments: These are self-reports and assessments made by the test participants. They are often collected via questionnaires filled out at the completion of the study that include structured rating scale questions as well as open-ended questions asking for short responses. Test participant assessments can also be collected via final verbal feedback debrief periods that typically occur at the completion of an evaluation.

Physiological Measures: There is recently great interest in using physiological measures such as heart rate and electric brain activity as an objective, unobtrusive way of assessing mental processing and workload. Eye tracking is also often classified under physiological measures (Matthews & Reinerman-Jones, 2017). In general, physiological measures tend to be highly indirect, and often noisy measures of macrocognitive functions but are included here as they represent current research frontiers.

Some of these measures are objective in the sense that the answer obtained is independent of who is making the measurement. For example, outcome performance measures (e.g., % correct and mean response time) can be assessed objectively provided that it is easy to determine what the correct response is. Other measures such as user assessments of the technology and expert judgments of performance are more subjective in that they depend on opinions and general impressions. However, objective measures are not necessarily always preferred over subjective measures. There are a number of other considerations in determining the appropriateness of a measure (Matthews & Reinerman-Jones, 2017). These include:

- ***Sensitivity*****:** Does the measure allow fine grained differences to be detected? For example, can it distinguish fine grained differences in level of workload?
- ***Diagnosticity*****:** Does the measure allow a detailed understanding of why a response was made and pinpoint the basis of performance problems? For example, if an error occurs, it is desirable to identify which macrocognitive functions contributed to the error. Is the error because: relevant information was not detected; or that the participant failed to correctly interpret and integrate the information into a correct understanding; or that they understood the situation correctly but formulated a wrong course of action; or that the course of action they came up with was correct, but it was poorly executed?
- ***Selectivity/validity*****:** Is the measure correctly getting at the thing you are interested in or is it measuring something different or more general? For example, does a heartrate measure provide a true indication of workload or is it measuring something else or something more general such as stress, fatigue, or arousal level?
- ***Reliability:*** Will the measure produce the same result each time a measurement (of the same thing) is taken? For example, will a person give the same workload rating each time they are presented with the same situation?
- ***Intrusiveness:*** Will the measure interfere with how the task of interest is performed? For example, if you interrupt someone as they are performing an ongoing task, will it distract them or otherwise change how they would otherwise perform the task?

As we will discuss below, in some cases subjective measures such as participant self-reports and feedback collected via post-study questionnaires have benefits over more objective measures with respect to sensitivity, diagnosticity, and selectivity/validity. In many cases the self-report measures have been shown to be highly reliable and are non-intrusive.

Below we present a range of measures that can be used to assess the different macrocognitive functions. We include outcome, process, test participant assessments, and physiological measures and discuss their relative merits. For each macrocognitive function we will also discuss methodological considerations when conducting evaluation studies addressing that macrocognitive function, and end with recommendations for successful evaluations. A summary of the measures that can be used to evaluate the impact of technology on each of the macrocognitive functions is provided in Table 2.

SENSEMAKING

Sensemaking is a core macrocognitive function that is responsible for a person's understanding of the situation they are in. It is what guides attention and drives intentions, plans, and actions. Sensemaking is closely related to Endsley's Level 2 SA (comprehension) in that sensemaking connects the various information that have been gathered into a meaningful understanding of the situation (Endsley, 1995a). Level 1 SA (perception of the elements in the environment) is an input to sensemaking. Level 3 SA (projection of future state) is an output of sensemaking, in that people anticipate future outcomes based on their current understanding of the situation.

There are many alternative methods that have been developed for measuring SA and sensemaking. Reviews can be found in Endsley (2021) as well as Tenney and Pew (2006).

Outcome Performance Measures

One approach is to simply measure outcome performance on an operational task of interest that to perform well depends on having correct SA and sensemaking. For example, Mosier and colleagues (2007) presented regional pilots with challenging aircraft malfunction scenarios (e.g., an engine is on fire). The measure of interest was whether they correctly diagnosed the malfunction and the time to respond. While correct outcome performance can be used to infer correct SA and sensemaking, the relationship is not necessarily one to one. For example, you can have a correct understanding of the situation but still take the wrong action (e.g., because of misunderstanding the rules of engagement). You can also take the right action in spite of having a completely wrong understanding of the situation. The inability to be diagnostic with respect to how the different macrocognitive functions contributed to a correct (or incorrect) response is a general limitation of outcome measures. Consequently, most researchers recommend collecting process measures that directly tap the content of SA as well as outcome measures.

Process Measures

There are multiple process measures of SA that have been developed. Most use measures that directly assess the content of SA - what information people have noticed (Level 1 SA), what they concluded about the situation they are in (Sensemaking/Level 2 SA), as well as their inferences about future events to expect (Level 3 SA). Direct measures of SA include:

- ***Event detection methods*** – Critical events are inserted in a scenario (e.g., an enemy target appears) and the participant is asked to respond when they detect it (Billman et al., 2020; Olmos et al., 2000; Gugerty & Falzetta, 2005).
- ***Embedded real-time probes*** – The participant is asked to articulate their current understanding of the situation. For example, a confederate (someone part of the study team) might pretend they are a commander and ask the pilot in the study to indicate their assessment of the situation and planned actions (Endsley, 2021).
- ***Think-aloud protocols*** – The participant is asked to 'think-aloud' as they are performing the task. In this way, which information they notice, and their assessment of the situation (sensemaking) can be directly assessed.
- ***The Situation Awareness Global Assessment Technique (SAGAT)*** – This technique involves freezing a scenario at multiple points in time, blanking out the screen, and asking individuals to answer Level 1 (perception), Level 2 (comprehension), and Level 3 (projection) questions based on memory (Endsley, 1995b).
- ***The Situation Present Assessment Method (SPAM)*** – This technique uses similar queries as SAGAT but does not require freezing the scenario. The queries are presented in real time while the participant is performing their task (Durso et al., 2004). To minimize intrusiveness, SPAM first provides a ready prompt. Once participants indicate they are ready, the SPAM queries are presented. The answers provided and response time (for those responses that are correct) are taken as indicators of the operator's SA. As an additional feature, the SPAM queries are treated as a secondary task. The time from when the 'ready' prompt appears and when the participants indicate they are ready becomes a real-time measure of the workload associated with the primary task they are performing. This has been used effectively to measure workload when evaluating the impact of novel displays on an air traffic control task (Trapsilawati et al., 2016).

Observation of behaviors and communications have also been used to infer SA. For example, expert observers may be used to evaluate what information has been noticed and what members of the team understood about the situation based on observing the behavior, communications, and actions among the team members (Tenney & Pew, 2006). In some cases, expert observers (e.g., experienced pilots) can detect problems in SA and sensemaking that are missed by other methods. For example, in the SUMIT program, researchers noted that there were many instances where expert observers identified that participants missed a threat or high value target passing through a sensor field of view because the participants were looking elsewhere, missed a chat message indicating an asset was taking fire, or missed a new threat icon appearing on a map (Alicia et al., 2020, page 27). They noted that these deficiencies in SA were not picked up by other measures of SA used such as SPAM and SART.

A final technique for uncovering what participants detected and understood, as well as the rationale for the actions they took is to conduct *a post-event debrief*. In a post-event debrief, participants are

asked to explain what happened during the scenario they just participated in (Billman et al., 2020). The *post event debrief methodology* is most appropriate when running more complex, high-fidelity simulations where there is a desire to avoid any artificial interruptions. Note that a post-event debrief that occurs immediately after a scenario is completed with the purpose of asking the participant to describe what they noticed, what they understood, and why they made the decisions they did while that information is still fresh in their minds. This is different from a *final verbal feedback debrief* that typically occurs at the completion of the evaluation study and focuses more on eliciting participant feedback on the technology, the problems they encountered, and opportunities for improvement.

Participant Assessments

Another approach to measuring SA is to rely on participant assessments of their own SA in a given situation. Participants are typically asked to rate their own SA using multi-point rating scale questions (e.g., 'How good is your awareness of the situation' on a scale from 1-7). These self-report measures of SA are considered to be subjective measures of SA in that they rely on participant judgments of their own SA. This contrasts with the various direct measures of SA that establish what the participant actually perceived and understood. The *Situation Awareness Rating Technique (SART)* is one of the first and most commonly used self-report measure of SA (Selcon & Taylor, 1990). It includes multiple rating questions covering three areas hypothesized to be relevant to SA: operator understanding of the situation, demands placed on attention, and available attentional resources. Endsley (2020) performed a meta-analysis comparing direct SA measures (e.g., SAGAT, SPAM) vs. self-report measures of SA such as SART. She concludes that self-report measures of SA tend to deviate from the results of SA measures that more directly measure the content of SA. Based on this analysis, she concluded that self-report measures are more useful at assessing people's confidence in their own SA rather than as a means to assess their actual SA.

Physiological Measures

Eye tracking has been used to assess SA because it is considered objective and unobtrusive; however, it is a highly indirect measure of SA (Endsley, 2021). While eye movements might tell you that the person looked at the relevant information, it doesn't indicate whether the information was in fact processed, correctly understood, or appropriately integrated into a coherent understanding of the situation (sensemaking).

There are currently explorations of use of other types of physiological measures to assess SA (Bracken et al., 2021). Bracken and colleagues argue that there is currently greater availability of portable sensor devices (e.g., neurophysiological monitoring devices) and fast computational algorithms that make it possible to explore the possibility of continuous, real-time assessment of SA in real-world settings. For example, one could use brain wave measures such as *electroencephalogram (EEG)* to assess low task engagement which would presumably imply low levels of SA. Similarly evoked brain potential measures such as *P300* could be used to assess when a given event triggers a high P300, suggesting surprise. Based on this, one could infer that SA Levels 2 and 3 were low since they did not anticipate the event resulting in a need for sensemaking activity to account for the new observations. While these potential applications of physiological measures are interesting to explore, they remain in their infancy. Further, they can only be used to infer whether the individual is actively engaged and whether they are surprised by events

occurring, the physiological measures do not provide any insight into the content of the person's SA (i.e., what they noticed and what they understood).

Sensemaking Methodological Considerations

Of foremost importance when assessing the impact of new technology on SA and sensemaking is to create test scenarios that allow aspects of SA and sensemaking to be externalized so that it can be directly observed and measured. This includes inserting specific events for participants to notice, orient to, comment on, and/or integrate into their ongoing understanding of the situation. Scenarios that include unexpected events that require participants to revise their understanding of a situation are particularly useful in assessing people's ongoing SA and sensemaking activity (Rankin et al., 2016; Landman et al., 2018). Including confederates in the study, individuals that are part of the experiment team that take on the role of a co-pilot or commander, also provides a natural way to encourage the test participant to externally verbalize what they have noticed, what they find surprising, their understanding of the situation, what they expect to happen next, and what actions they believe should be taken as a consequence.

DIRECTING ATTENTION/MANAGING ATTENTION

Directing attention and managing attention to focus on the most critical tasks are essential activities for pilots, and a cornerstone of the macrocognitive function. Directing attention refers to where an individual chooses to focus their attention and whether they are attending to the right task. Managing attention refers to deciding whether and when to shift attention across multiple competing demands for attention (e.g., whether and when to interrupt a given task so as respond to a request from a person or a prompt for attention from a technology). Often, pilots are inundated with data and information; the challenge is to determine what to focus on, what to prioritize, and what to ignore temporarily. Tools to support pilots must not interfere with their ability to manage their attention (e.g., through constant interruption) and should ideally help direct pilots' attention to critical information.

Directing Attention

Outcome Performance Measures

One approach to evaluating attention focus is to assess outcome performance – did the operator correctly accomplish the task and how long did it take? Outcome performance measures include *accuracy and detection measures* (Hameed et al., 2009), and *response time measures* (Suh & Ferris, 2019; Schriver et al., 2008). The assumption is that if participants performed the task correctly then they must have correctly detected and understood the relevant information. However, as discussed in the section on sensemaking, this inference is not necessarily valid.

Process Measures

Another way to measure focus of attention is to assess what information they detected. This is essentially what measures of SA level 1 are designed to answer. Relevant measures of SA are summarized in the section on sensemaking and include event detection methods, SAGAT (Endsley 1988), and SPAM (Durso & Dattel, 2004). Because these techniques assess what elements in the environment people have detected, they allow attention to be understood in the moment and can be thought of as direct measures of attention.

Physiological Measures

Eye tracking is also used to assess focus of attention. Eye track measures include *gaze direction, dwell time,* and *progression of fixations*. This choice assumes that gaze reveals attention. In research specifically focused on assessing attention, Kinney and O'Hare (2020) and Mumaw, Billman, and Feary (2020) use percent dwell time as the best way to assess attention. In contrast, others look at movements to areas of interest (Moacdieh et al., 2020; Ophir-Arbelle et al., 2013; Grundgeiger et al., 2022). Another approach to using eye tracking is to focus on *relevant vs. irrelevant fixations*, comparing them before and after the target appears (Ratwani & Trafton 2010; Foroughi et al., 2021; Vogelpohl et al., 2020).

Other physiological measures such as *pupil dilation* (Kinney & O'Hare, 2020) and *heart rate variability* (Berry et al., 2021; Kinney & O'Hare, 2020) can give researchers a window into shifts in attention that may cue a change far before conscious knowledge has occurred. A challenge with these is that they do not provide insight into where attention is shifting or why, but these can be excellent companion measures to more explicit probes.

Managing Attention

When new pilot aiding technology is introduced in the cockpit it important to understand what impact it may have on pilot attention management. Questions include:

- Does the system interrupt at appropriate times so as to avoid impacting other workflows?
- Is the system effective at drawing attention to the right information?
- How does the system support the pilot in refocusing attention on the original task when they get interrupted for whatever reason (either the system itself or external interruptions)?

These types of questions are typically answered using process measures that examine responses to interruptions.

Process Measures

Many of the measures of managing attention use interruptions to assess how a participant switches tasks and then resumes the original task again. For example, researchers look at the lag time to transition to the interrupting task as well as lag time to resume the original task once the interrupting task is completed (e.g., Berry et al., 2021). Researchers also measure the impact of interruptions on the accuracy of performing the original (primary) task when a secondary task

interrupts (Ratwani & Trafton 2010; Lu et al., 2013). In a study assessing driving behavior, researchers used an algorithm that predicted attention levels (Semmens et al. 2019). Semmens and colleagues measured attention in the context of an actual driving task, focusing on attention during high and low complexity contexts. As participants were driving, they would be asked "is now a good time?" to interrupt, researchers used lag time to respond as a way of inferring attention availability.

Directing/Managing Attention Methodological Considerations

Many articles recommend examining attention focus and attention management under both high and lower workload conditions to see how the tools aid pilots when attention is taxed, but also when it might be easy for attention to wander. Under high workload conditions when there are many potential distractors, pilots are forced to decide what to attend to and when. Conversely, when there is little going on, they may lose focus and miss new cues.

Landman and colleagues (2017, 2018) suggest including unexpected events to see how pilots react and how tools can support directing attention appropriately. Billman and colleagues (2021) also recommend using a confederate to introduce errors so that researchers can assess how attention is managed when information is unclear or conflicting.

MANAGING WORKLOAD

Pilot workload is an important factor to consider when evaluating the effects of technology in decision making and SA. While the term "workload" can refer to physical and cognitive activities, this section focuses on ways to measure cognitive workload. These measures assume a resource-demand framework for characterizing workload – cognitive resources are limited, thus when demand exceeds resource availability, overload is experienced (Vogl et al., 2021). Because cognitive overload can be caused by a variety of factors other than task-related workload, researchers recognize the criticality of using a combination of measures to evaluate workload.

Outcome Performance Measures

One way to evaluate workload is to measure *task performance*, usually in terms of time to complete, number of errors, etc. (Ernst et al., 2020). Performance on the primary task of interest can be measured, but it might require all of a pilot's mental capacity to perform satisfactorily, leaving no spare capacity to handle any additional tasks might arise.

Performance on a secondary task can be used as an indicator of spare cognitive capacity. Measuring secondary task performance involves measuring timeliness and accuracy of a secondary task while an operator is completing a primary task which is the higher priority task. An example of a secondary task to measure includes measuring a pilot's responses to verbal communication prompts while flying. Performance on a secondary task is expected to deteriorate as the primary task demands increase.

Participant Assessments

Workload is also measured using self-report measures. Commonly used questionnaires include the *NASA-TLX* (Falkland et al., 2020; Loft et al., 2015; Hart & Staveland, 1988), the *Subjective Workload Assessment Technique* (SWAT), the *Bedford Workload Rating Scale* (Roscoe & Ellis, 1990), and *Continuous Subjective Workload Assessment Graph* (C-SWAG; Gawron, 2019; Berry et al., 2021). Interestingly, the SPAM method for assessing SA, is also used to measure workload by serving as a secondary task. SPAM introduces prompts designed to evaluate operators' SA at intervals throughout an experimental task. A 'ready' prompt appears, and the test participant is asked to press a key when they are ready to start to answer the SA questions. The time from when the 'ready' prompt appears and when the participant presses the key to indicate they are ready for the SA questions becomes a real-time measure of the workload associated with the primary task they are performing (Trapsilawati et al., 2016; Trapsilawati et al., 2017). If workload is low, operators can respond quickly to the ready prompt. If workload on the primary task is high, operators will take longer to press the key indicating they are ready for the SA queries.

Physiological Measures

Of physiological measures, cardiovascular measures (heart rate, heart rate variability, etc.), EEG, and *electrocardiogram (ECG)* are most strongly correlated with workload. Cardiovascular measures, such as heart rate, heart rate variability, heart period, and blood pressure have all been correlated with other measures of workload (Hughes et al., 2019). ECG has also been correlated with other measures of workload (Martin et al., 2019). Ocular indices, such as pupil diameter, fixation duration, and fixation count have shown mixed results in relation to workload (Bhaskara et al., 2020; Mercado et al., 2016). EEG has been used to track changes in workload and SA in air traffic controllers (Trapsilawati et al., 2020). EEG data indicates types of brain activity and location of that activity involved in high workload and high stress situations (Trapsilawati et al., 2020), which can be used to infer the level of workload being experienced by the participant.

Because workload overlaps with other physiological processes, such as emotion, stress, and wakefulness, physiological measures of workload can be affected by many factors, making it difficult to tease out the unique influences of task-related workload (Vogl et al., 2021; Matthews & Reinerman-Jones, 2017). Further, because physiological data should be compared to participants' baseline data rather than directly to other participants, it requires substantial pre-processing (Zhang et al., 2020) and can be difficult to interpret and analyze (Matthews & Reinerman-Jones, 2017). In addition, brain imaging techniques such as EEG can be difficult to employ because workload is not associated with specific areas in the brain; it seems to be an emergent property related to attention (Matthews & Reinerman-Jones, 2017).

While eye tracking is another potential measure of workload, it is a less reliable indicator (Martin et al., 2019; Matthews & Reinerman-Jones, 2017). Eye tracking may be a better indicator of attention allocation rather than workload per se.

Managing Workload Methodological Considerations

Many researchers suggest combining several different types of measures when evaluating workload (Schnell et al., 2008; Hughes et al., 2019; Martin et al., 2019; Matthews & Reinerman-

Jones, 2017; Vogl et al., 2021; Trapsilawati et al., 2020). Vogl and colleagues (2021) recommend a composite measure of workload that combines performance, physiological, and subjective measures.

PLANNING

In military aviation, pilots routinely engage in planning as well as dynamic replanning activities. For example, a pilot may need to plan the flight route to get from their point of origin to a particular destination avoiding known threats and obstacles as part of the mission planning process. Another common military aviation planning task is allocating available resources to accomplish mission objectives. For example, in a FVL reconnaissance mission, the air mission commander (AMC) may need to decide how to allocate the available manned and unmanned resources (e.g., ALEs) to provide appropriate coverage of the reconnaissance area.

In dynamic environments, planning is rarely a once and for all activity. Rather, people are constantly assessing whether their current plans are appropriate to their evolving understanding of the situation and goals and revising the plan accordingly. In many cases this requires *adapting the plan* or even coming up with an entirely new plan on the fly. In the case of route planning, if unanticipated threats or obstacles are detected in route, or if the mission gets redirected, then the pilot will need to dynamically revise the route to accommodate the changing situation. Similarly, in the case of resource allocation plans, if the situation changes (e.g., an ALE malfunctions or is shot down; or a new target is detected) then the AMC will need to rapidly revise the asset allocation plan to accommodate changing capabilities and priorities.

Currently there are a number of new technologies being developed to support pilot planning and dynamic replanning. These include new kinds of visualizations and decision aids for route planning, obstacle detection, and dynamic route replanning (e.g., Friesen, Borst, Pavel & Stroosma, 2021; Friesen, Borst, Pavel, Masarati & Mulder, 2021; Szoboszlay et al., 2021) as well as automated systems to support managing multiple unmanned systems including asset allocation and dynamic reallocation to achieve mission objectives such as a base defense task (e.g., Alicia et al., 2020; Calhoun, Bartik, et al., 2020; Stowers et al., 2020).

There are multiple methods and measures available to assess the planning process and the impact of new technologies on planning. The most straightforward involve measures of outcome performance – that is how well the task was performed.

Outcome Performance Measures

Outcome performance measures include the time required to complete the planning task (i.e., response time), and the quality of the plan generated, and actions taken. Plan quality include whether the plan adequately took into account the mission constraints (e.g., for a flight routing task this might include threat location, flight restrictions, terrain, and vertical obstacles) as well as how effective the plan was (e.g., did the route plan minimize distance travelled and fuel used? Did the asset allocation plan allocate the closest asset that was available and could achieve the task assigned?).

There are numerous specific examples of outcome performance measures used in military planning tasks. In this case, rather than assessing the quality of the plan directly, actions taken during execution are considered a representation of the plan. In a study examining the impact of alternative head-up displays on helicopter pilot obstacle avoidance during flight, Friesen, Borst, Pavel, Stroosma, Masarati and Multer (2021) included flight control measures such as deviation from ideal flight altitude, lateral position and speed, as well as measures of safety margin (e.g., the vertical clearance of the climb-over maneuvers over obstacles). The SUMIT program examined a variety of outcome variables such as time to complete areas and route reconnaissance tasks, sensor utilization efficiency, duration of friendly assets vulnerable to threats; fraction of the ingress route and landing zone scanned by a sensor; scanning additional areas of interest that emerge, locating a high value target, and destroying an anti-aircraft weapon system.

Another important outcome measure relates to *resilience* in the face of unforeseen conditions, particularly conditions that are beyond the competence envelope of the aiding technology (Woods, 2015). The outcome performance measure of interest is the ability of the joint person-technology system to dynamically adapt plans so as to continue to meet mission objectives. Hoffman and Hancock (2017) discuss multiple *measures of resilience*. One approach is to create complex situations that stress the boundaries of competence of the automated planning aid so that it fails or generates poor solutions, and measure how the person-technology system jointly handle the situation. Is the pilot able to detect that the situation is beyond the competence of the aiding technology? Are they able to manually take over and perform successfully? Are they able to redirect it (e.g., by providing information it may not be aware of, or placing constraints on the kind of solution it generates) so that it generates a useful planning solution? Are they able to leverage some of its outputs (e.g., its displays) to support them in coming up with a good planning solution themselves? Among the measures that can be used to assess resilience include:

- Ability of the pilot to recognize that the situation is beyond the competence envelope of the aiding technology,
- Quality of the planning solution generated,
- Time to generate the solution.

A variant is to compare the performance of pilots using the aid to their performance without the aid. The question of interest is whether outcome performance with the aid is better than performance without the aid in both routine situations for which the aid is designed to handle as well as situations beyond the capabilities of the planning aid.

Process Measures

In addition to outcome measures, researchers typically also collect process measures. These include what information and consideration the individual took into account in generating the plan. This applies both in the case of routine planning situations as well as situations that require dynamic adaptive replanning. In the case of unanticipated conditions, process measures of interest include: did they recognize that the situation deviated from expectations? Did this trigger sensemaking activity and what was their revised understanding of the situation? Did they recognize a change in goals and priorities, and what factors did they consider in revising the plan?

Among the types of process measures that can be used to address these types of questions include:

- Embedded real-time probes,
- Think aloud,
- Post-event debriefs.

These process measures are described in detail in the section on Sensemaking.

Planning methodological considerations

In examining the impact of an aiding technology on pilot planning and replanning, it is important to include a range of scenarios, including situations that are beyond the competence boundaries of the aiding technology. Including scenarios where errors or design breakdowns are intentionally introduced, assures that adaptive response will be required (Hoffman & Hancock, 2017). It is also important to include process measures as well as outcome measures so as to be able to understand the information the participants were aware of, the sensemaking activities triggered, what they understood, and what factors they considered in generating (or revising) the plan.

COMMUNICATING/COORDINATING (AMONG DISTRIBUTED TEAMS OF PEOPLE AND AUTOMATED AGENTS)

The introduction of new technology can change how teams of people function. New technologies can facilitate or disrupt team communication, coordination, and SA. These changes should be measured and evaluated to determine how new technologies impact the functioning of teams. Team constructs can be evaluated by combining individual measures, but there is a push to study team cognition at the team level by measuring team processes such as communication and coordination, and team states such as Team SA and Shared SA (Cooke & Gorman, 2009; McNeese et al., 2021; Huang et al., 2020).

Outcome Performance Measures

Outcomes team performance measures are variations on individual performance measures. Cooke and Gorman (2009) describe using perturbations, or *roadblocks*, during team-based simulations. Whether the team is able to overcome the roadblock (e.g., a disruption in communications) to achieve the scenario objective is one outcome measure. McNeese and colleagues (2021) take this a step further, describing the proportion of roadblocks overcome per mission as an outcome measure of Team SA. Mathieu and colleagues (2000) measured team performance by awarding points based on the objectives of a simulated mission (3 points if the team survives the mission, 2 points for each waypoint reached, 1 point per enemy plane shot down). The mission objectives are directly at odds with each other, so a team is required to coordinate and negotiate to try to achieve the highest score (*total mission score*).

Process Measures

Teamwork process measures include evaluating team communication and coordination. Communication among team members can be evaluated in terms of content (what is

communicated) and flow (who talks to whom, and when). Cooke and Gorman (2009) use communication-based event data analysis to quantify and evaluate communication and coordination in small teams. Coordination among team members can be evaluated in terms of synchronization of activities. Cooke and Gorman (2009) created a ratio score that could quantify team coordination over time by tracking communication flow with time onset of specific communication events during a simulation. Introducing perturbations into the simulation allowed them to quantify team coordination in response to the roadblock. Mathieu and colleagues (2000) used trained observers to rate team performance in terms of strategy formation and coordination, cooperation, and communication.

Another important process measure relates to SA at the team level. "Team SA" is a term that refers to the SA of each team member, whereas "Shared SA" is a term that refers to overlapping SA among team members (i.e., common ground; Huang et al., 2020). It is important to measure the effects of new technology on both team and shared SA. Individual measures of SA (e.g., SAGAT, SPAM) can measure shared SA by evaluating the amount of overlap between different team members' responses.

A measure of team SA, called the *Coordinated Awareness of Situation by Teams (CAST)* method, also uses perturbations to measure resulting team members' activity and interactions. CAST is used to score team activities compared to the true state of the world in terms of hits (optimal responses) and false alarms (suboptimal responses) (Cooke & Gorman, 2009; McNeese et al., 2021). The *Incorrect SA in Failed Team Tasks (iSAFT)* method (Huang et al., 2021) uses failed team tasks to identify what operators did not know that they should have known to articulate failures in team SA. Mathieu and colleagues (2000) measured task and team mental models using individual team members' ratings of relationships between attributes relevant to the team's goals. Rating matrices were analyzed using a network-analysis program to calculate a convergence score for mental models among the team.

Communicating/Coordinating Among Teams of People and Automated Agents: Methodological Considerations

Team effectiveness should be evaluated using a combination of measures, including team interaction processes (communication, coordination) and team states (team SA, team trust, team resilience, etc.) along with outcomes (Huang et al., 2020). Furthermore, simulation scenarios should present dilemmas and perturbations to measure how well teams are able to cope with unexpected events (Mathieu et al., 2000; Cooke & Gorman, 2009).

COMMUNICATING/COORDINATING (WITH TECHNOLOGICAL AIDS)

"The project of building interactive machines has more to gain by understanding the differences between human interaction and machine operation, than by simply assuming their similarity." - *Lucy Suchman*

Lucy Suchman (1987) so eloquently described the challenges of human-technology teaming: Technology should complement human skills, so performance is amplified. For the challenges discussed in this review, macrocognitive principles must be understood not just as they apply to individual pilots, but also to how pilots communicate and coordinate with the technology. There are three key facets of this dynamic that should be measured to understand how successful it is:

- How well trust in the technology aid can be built,
- How easy it is for pilots to understand what the technology aid is doing, and how successful the technology is at conveying its state and operations. This also encompasses the development of a mental model of the technology that allows the pilot to deeply understand and predict its behavior,
- How easy it is for the pilot to direct, and redirect, technology when appropriate.

In this section we discuss methods for measuring these three facets of human-machine teaming.

Trust

There has been extensive research on trust, both between people, and between humans and technology. Research on trust between humans and technology is most relevant here. Two primary methods that researchers have relied on are performance-based measures that assess whether and when users will rely on the technology and self-report measures of trust in the technology.

Outcome Performance Measures

Performance-based measures of trust include measuring reliance and compliance behaviors, then inferring the level of trust from these behaviors. Compliance is defined as when an operator takes an action consistent with what the automation presents (e.g., an operator takes an action in response to a warning system's alert; Meyer, 2004). Reliance is defined as when an operator does not initiate an action without receiving a prompt from the automation (e.g., an operator does not take an action when not alerted by an automated warning system; Meyer, 2004). Both behaviors are indicators of operator trust in the automation.

Automated systems may not be 100% reliable, but that does not necessarily need to negatively affect operators' trust in the system. Calibrated trust reflects the coherence between the automation's capabilities and the operator's trust in that automation (Lee & See, 2004). Signal Detection Theory has been applied to measure trust in automated systems. Calculating proper acceptance, correct rejections (hits and correct rejects) compared to disuse (misses) and misuse (false alarms) provides insight into how well an operator's trust is calibrated with the capabilities of the automation (Mercado et al., 2016; Stowers et al., 2020). Trapsilawati and colleagues (2016; 2021) used aid utilization rate as a measure of trust in a conflict resolution advisory aid for air traffic control. Aid utilization rate was defined as the ratio of accepted advisories relative to the total number of advisories provided by the conflict resolution advisory aid.

Participant Assessments

There are multiple self-report measures, many of which were based on models of interpersonal trust. These scales mostly differ in how they operationalize trust, how many factors are included, and type of relationship being evaluated (e.g., human-human, human-automation, human-robot, human-artificial intelligence, etc.; see Hoffman, Mueller, Klein, G., & Litman, 2018a for a nice summary of trust scales). Jian and colleagues (2000) created the *Checklist for Trust between People and Automation*, which defined trust as a trait with six factors (fidelity, loyalty, reliability, security, integrity, and familiarity). This is one of the earliest and most widely used rating scales of trust. The *Cahour-Forzy* (2009) scale defines trust in a cruise control system for cars in terms of three factors: reliability, predictability, and efficiency. Hoffman and colleagues (2018a) used the Cahour-Fourzy scale as a basis for their own scale to measure trust in explainable artificial intelligence (XAI), with questions to address whether the XAI system is predictable, reliable, efficient, and believable. The *Merritt* scale (2011) describes trust in automation as an emotional, attitudinal judgement based on five factors (belief, confidence, dependability, propensity to trust, and liking). The *Schaefer* scale (2013) measures machine performance and team collaboration as indicators of trust in human-robot collaboration. The SUMIT program used the Schaefer Trust in automation scale (Alicia et al., 2020).

Lyons and colleagues (2011) found that trust in automation includes orthogonal constructs of trust and distrust, each of which uniquely predicted decision confidence when using an automated decision aid. They suggest using separate scales of trust and distrust.

In many cases, researchers have developed their own versions of self-report rating scales that are more tailored to a specific application. For example, Mercado and colleagues (2016) used a modified version of the Jian and colleagues (2000) trust rating scale where they focused on two dimensions: (a) information analysis (trust in the information and analysis displayed) and (b) decision and action selection (trust in the suggestions and decisions presented). A trust rating scale incorporating these same two dimensions was also used by Stowers and colleagues (2020) in evaluating the IMPACT support system for managing a team of heterogenous unmanned vehicles in an Army base defense application.

Understandability/Explainability

Another key aspect of human-machine communication success is how well the humans understand how the system works so they can use systems with confidence – knowing when to trust and use the system as well as when not to follow its guidance when it is wrong. The DARPA XAI project extensively reviewed methods and measures of how to assess (and build) systems that are good partners (Hoffman, Meuller, Klein, G., & Litman, 2018b). The XAI research team recommended a variety of measures, including outcome measures of how much people rely on the systems and how reliable the systems are. These outcome measures are covered in the section on objective measures of trust. The XAI research team also developed a number of innovative process and participant assessment measures to evaluate how understandable and predictable the technologies are.

Process Measures

When exploring how understandable a complex system is, researchers stress the importance of using methods that help elicit the pilots' mental models of a system (Hoffman, Mueller, Klein, Litman, 2018b, Hoffman, Klein & Borders, 2018; Klein, Borders, Hoffman, & Mueller, 2021). Hoffman and colleagues (2018b) surveyed numerous potential techniques for understanding mental models including think-aloud protocols, question answering/structured interviews, self-explanation tasks, and prediction tasks that ask people to predict what an artificial intelligence system will do in various situations. One example is a post-study mental model questionnaire that asks the participant to provide short answers to a set of pointed questions about their mental model of how the system works. Example questions include, 'what features is it paying attention to?', 'what types of situations is it good at handling?', 'what types of situations is it poor at handling?', 'did the system ever surprise you?', as well as 'what factored into your decision of whether to accept or reject the system's recommendation?' One more recent measure of mental models is the *Mental Model Matrix* (Klein, Borders, Hoffman, & Mueller, 2021). This approach focuses on four distinct elements of their mental model of the system: how the system works, how the system fails, strategies the person has developed to work around the system or get it to work, as well as the kinds of errors the person themself was prone to making.

Another approach to elicit participant mental models of a technology is the *cued-retrospective method* (Hoffman et al., 2018b). In this method, study participants are shown replays of their interaction with the technology that occurred during the study and asked to explain and comment on these interactions, their understanding of how the technology worked, and their assessment of the technology. This approach was employed by Miller, Godfroy-Cooper and Szoboszlay (2021) in evaluating a novel obstacle cueing display for helicopter pilots as part of the Army's DVE-M program. An obstacle cueing-specific debrief was held within a few days of the test flight where pilots were asked for detailed impressions of the obstacle display, including features and modifications desired in future display development iterations.

Participant Assessments

The XAI research team also developed participant assessment measures that include a measure of overall communication effectiveness of the technology referred to as *communication scorecard* (Klein, Hoffman, & Mueller, 2021).

With the recent emphasis on generating systems that are transparent and explainable, measures of explainability are also being developed. Hoffman and colleagues (2018b) present a questionnaire that can be used to measure a person's assessment of *'explanation satisfaction'*. Explanation satisfaction is defined as the degree to which the person feels they understand the system being explained.

Directing the Technology/Directability

Another key element to evaluating the success of human-machine communication is *directability* - how easy it is for the pilots to direct and redirect the technology as needs or contexts shift. Wiggins and Cox (2010) discuss several methods to assess directability, including both objective

outcome performance measures as well as participant assessment measures. For outcome performance measures, they suggest using a scenario where the system gives an incorrect answer, and the pilot must redirect it. Success is measured in both performance accuracy and speed of redirect. An example of a participant assessment method they recommend is a *post-session questionnaire* to obtain participant feedback on how easy it is to modify/redirect the system when appropriate.

Overall Assessment of Technology Effectiveness

It is also important to obtain participant assessment of the overall useability and usefulness of the technology in supporting their work (Roth, Bisantz, et al., 2021). Approaches for eliciting study participant assessment of the technology include conducting a *final verbal feedback debrief* session following the more formal evaluation. In the final verbal feedback debrief structured interview questions are used to elicit participant assessment of the technology. This approach was used in the SUMIT program to obtain participant feedback on the strengths and limitations of the technologies being evaluated (Alicia, et al., 2020).

Post study *user-feedback questionnaires* can be used as an alternative to or complement to final verbal debriefs (Roth, Bisantz, et al., 2021). The study questionnaire can include closed-form rating questions as well as more open form questions that require participants to write a sentence or two explaining aspects of the system they thought were most helpful, aspects that they thought were suboptimal, and recommendations for improvement. Questions should probe how easy it was to understand and evaluate the system's behavior as well as how easy it was to redirect the technology in situations where the recommendations it provided were incorrect or suboptimal.

Communicating/Coordinating with Technology Aids: Methodological Considerations

There are important methodological considerations in designing evaluations to assess different aspects of person-technology communication and coordination. For example, there are several key considerations to keep in mind when evaluating trust in a new technology. Chiou and Lee (2021) suggest that experiments should manipulate trust in some way (e.g., by manipulating the reliability of the system, alter trust-signaling behaviors, and creating automation failures) then measuring the outcomes of these manipulations on subjective measures of trust. Kaplan and colleagues (2021) found that shared risk is an important contextual factor that influences trust. Scenarios and simulations should include risky situations. Trust is not a static state that is achieved at the end of an experiment; rather, it is a dynamic process that is influenced by many factors throughout a scenario. Hoffman and colleagues (2018a) suggest integrating trust metrics throughout a test scenario to generate a more valid picture of operator trust in an automated system. Hoffman, Johnson, and colleagues (2013) describe how trust is partially dependent on past experiences, thus evaluations of new automated systems should include edge cases and contexts in which trust in the automation is inappropriate. Moreover, participants in such evaluations should receive training in how the system operates so they are familiar with it before trust is evaluated. Finally, because artificial intelligence behaviors can evolve over time, automation that incorporates artificial intelligence should be re-evaluated in terms of its operations and operator trust over the course of its use and re-standardized at regular intervals (Kaplan et al., 2021).

Table 2 Measures for assessing the impact of new technologies on macrocognitive functions.

Macrocognitive Functions	Outcome Performance Measures	Process Measures	Test Participant Assessments	Physiological Measures
Sensemaking (includes situation awareness)	• Quality of performance on operational task (e.g., % correct) • Time to complete operational task	• Event detection methods • Embedded real-time probes • Think aloud protocols • SAGAT • SPAM • Observations of behaviors and communications • Post event debriefs	• SART • Multi-point rating scale questions in post-test questionnaires	• EEG • P300 • Eye Movements
Directing attention (includes managing attention)	• Detection time • Detection Accuracy	• Measures of Level 1 SA (e.g., event detection methods, SPAM, SAGAT) • Interruption tasks (lag time to transition and resume, accuracy)		• Eye tracking fixations (durations, count, relevance, retrospective vs. prospective) • Pupil dilation • Heart rate variability
Managing workload	• Primary task performance • Secondary task performance (including SPAM as secondary task)		• NASA-TLX • Bedford Workload Scale • SWAT • C-SWAG	• Cardiac measures (heartrate variability, heart rate, heart period, blood pressure, ECG) • EEG • Ocular indices of pupil diameter, fixation duration, and fixation count
Planning	• Quality of plan generated • Time to generate plan • Measures of resilience	• Embedded real-time probes • Think aloud protocols • Post event debriefs	• Multi-point rating scale questions in post-test questionnaires	

Table 2 continued

Macrocognitive Functions	Outcome Performance Measures	Process Measures	Test Participant Assessments	Physiological Measures
Communicating (with teams of people and automated agents)	• Achieve scenario objective • Proportion of roadblocks overcome per mission • Total mission score	• Communication content • Communication flow event data analysis • Ratio score of team coordination • Observer ratings of team processes • Measures of shared SA (e.g., overlap between individuals' SAGAT, SPAM responses, convergence of team members' mental models) • CAST • iSAFT		
Communicating (with technology aids)	• Measures of reliance and compliance • Signal detection theory-based measures of calibrated trust • Measures of utilization rate • Measures of directability (speed and accuracy)	• Think-aloud protocols • Question answering/structured interview • Prediction tasks • Mental model questionnaires • Mental model matrix task • Cued retrospective protocols	• Self-report trust scales • XAI communication scorecard • XAI explanation satisfaction measures • Participant assessments of directability • User feedback questionnaires • Final verbal feedback debrief	

SUMMARY OF METHODOLOGICAL RECOMMENDATIONS

Beyond capturing the measures to use in evaluating new technology, it is also critical to explore the methods and strategies for structuring evaluations. It is important to match the type of rigor to the evaluation goals. For basic research, rigor often involves isolating a small set of variables and controlling others to understand the contribution of individual components. For applied research, it is more important to understand how decision making happens in a realistic context. In this case, a more rigorous design focuses on ecological validity, in which the test environment maps onto real-world complexities. Full scope simulators allow for a blend of control and realism. The goals of the study guide which variables to control, and which real-world complexities to include.

Designing evaluation scenarios involves making trade-offs: what are the goals for the research, what is critical to carefully control, and what is vital to be true-to-real-life (see many chapters in Patterson & Miller, 2010)? G. Klein and his colleagues' caution against assuming that every dimension should be carefully controlled when evaluating AI systems —the more control, the less useful the research is in evaluating how a system will be used in real life (Klein, Jalaeian, Hoffman, Mueller, & Clancey, 2021). Evaluating a system solely in a series of tightly controlled laboratory settings with undergraduate students might result in lots of insights, but they might not be transferrable to the high-pressure world of an expert pilot. In contrast, only doing research in the field might have too much variability to yield systematic and useful results.

One recommendation is to use different methods at different points of the system development lifecycle, with the guiding mindset to aim for "Minimum Necessary Rigor" (Klein, G. et al, 2021) to only evaluate the most critical elements, as efficiently as possible. Klein and colleagues also recommend smaller, iterative studies; limiting research to just a few conditions; and focusing on how the systems help or hinder performance, rather than more abstract research questions.

With that spirit in mind, there are also some more specific recommendations relevant to designing good evaluation studies (and see Roth, Klein, & Ernst, 2021 for more details).

Evaluation happens best in scenarios: vignettes of a situation where the pilot must engage in macrocognitive functions to make decisions. Good scenarios feel true to life, challenging, and engaging.

- Create scenarios that specifically challenge the relevant macrocognitive functions being studied. If a new tool is intended to support sensemaking, craft a scenario that forces the pilot to figure out what is going on in a confusing situation with conflicting information.
- Ideally include a range of scenarios that vary in complexity. These include straightforward cases that should be easy to handle, more cognitively challenging situations where the aiding technology would be expected to be particularly beneficial, and 'edge cases' that are likely to be beyond the capabilities of the technology. Edge cases provide a 'stress test' and can include situations where the technology performs sub-optimally or entirely fails. They are included to evaluate the resilience of the joint person-technology system. They address the question 'what

happens in situations where the technology is beyond its competence envelope'? Is the pilot able to recognize these situations and effectively compensate?

The evaluation study should ideally compare performance in two or more conditions so as to better understand the impact of the new technology on performance. Examples of potentially relevant comparison conditions include pilot performance with or without use of the new technology, when the technology is reliable a 100% of the time vs. when it fails a certain percentage of the time or for certain types of situations, with experts versus with less experienced pilots. It is particularly useful to compare performance with and without the technology aid to assess under which conditions the aiding technology is particularly useful in improving performance, as well as whether there are any conditions under which performance with the technology aid is worse than without it (e.g., instances of automation bias).

Use multiple complementary measures whenever possible, to include outcome measures, process measures, as well as participant assessment measures. While outcome measures are arguably the most relevant from an operational perspective, they are not as sensitive or diagnostic as process measures. For example, if performance is good but workload is very high, it may mean that the pilot is working too hard to maintain good performance and would be unable to handle any additional challenges. Participant assessments are also important because they provide complementary information that cannot otherwise be obtained. For example, participant feedback, elicited via post-study questionnaires and/or post study final verbal feedback debriefs, on how the technology operated and suggestions for improvement can be invaluable in insuring that the final implemented system will be usable, useful, and robust.

Physiological measures should generally be coupled with other measures (e.g., process measures, outcome measures and/or participant assessment measures) because of the current state of their maturity. While progress is being made in developing robust suites of physiological measures for dynamic real time measurement (e.g., Schnell et al., 2008), for the most part physiological measures have not been shown to be sufficiently selective, and diagnostic to be relied on as the sole measure (Matthews & Reinerman-Jones, 2017). For example, heart rate variability, while a potential measure of mental workload is also associated with emotion regulation and psychological fatigue (Mathews & Reinerman-Jones, 2017) indicating a problem in *selectivity*. Similarly, while evoked brain potential P300 signals surprise suggesting a potential problem in SA or sensemaking, it doesn't provide any insights into what the person didn't know or understand, indicating a limitation in *diagnosticity*.

Create an iterative research plan designed to learn and adapt before the next test. Try to avoid the "everything and the kitchen sink" exhaustive and complex study design.

Aviation tasks often involve interaction across multi-person teams, and it is important to evaluate the impact of new technology not just on individual performance but on team performance and team dynamics.

7 Pulling it All Together

In the previous chapter we summarized a variety of methods and measures available to evaluate new technologies that are intended to support pilot SA and decision making, and provided recommendations for study design and measurement selection.

In this chapter, we pull it all together by illustrating our recommendations for design of evaluation studies and selection of evaluation measures using a concrete example. In this example, we are evaluating a hypothetical landing aid intended to support the pilot flying in a FVL context. We begin with an overview of the various considerations that inform the design of an evaluation study and the selection of evaluation measures. We then walk through the illustrative example, showing the kinds of specific considerations that might arise and how those would be used to inform the design of the study, selection of evaluation test bed, and selection of evaluation measures.

In the Appendix, we provide two additional evaluation examples for a different FVL aiding technology -- a hypothetical ALE management system intended to support an AMC managing multiple ALEs.

CONSIDERATIONS IN DESIGN OF TECHNOLOGY EVALUATION STUDIES

Figure 4 provides a high-level description of various considerations that go into design of an evaluation study and selection of measures (Roth & Eggleston, 2018; Roth, Bisantz et al., 2021). We first describe these considerations and then illustrate their application in design of an evaluation study, including selection of measures for the hypothetical technology aid for FVL pilot landing.

Design of an evaluation starts by identifying the ***evaluation questions***. These are the specific questions to be answered as part of the evaluation study. They are sometimes referred to as the hypotheses to be tested. The evaluation questions are typically derived from three considerations:

- **Hypotheses of support:** These are the hypothesized performance benefits anticipated by system developers, and organizational leadership. Ideally the hypotheses of support should be framed in terms of the anticipated impact on the different macrocognitive functions. Examples include reducing the decision cycle time, improving SA, and reducing workload.

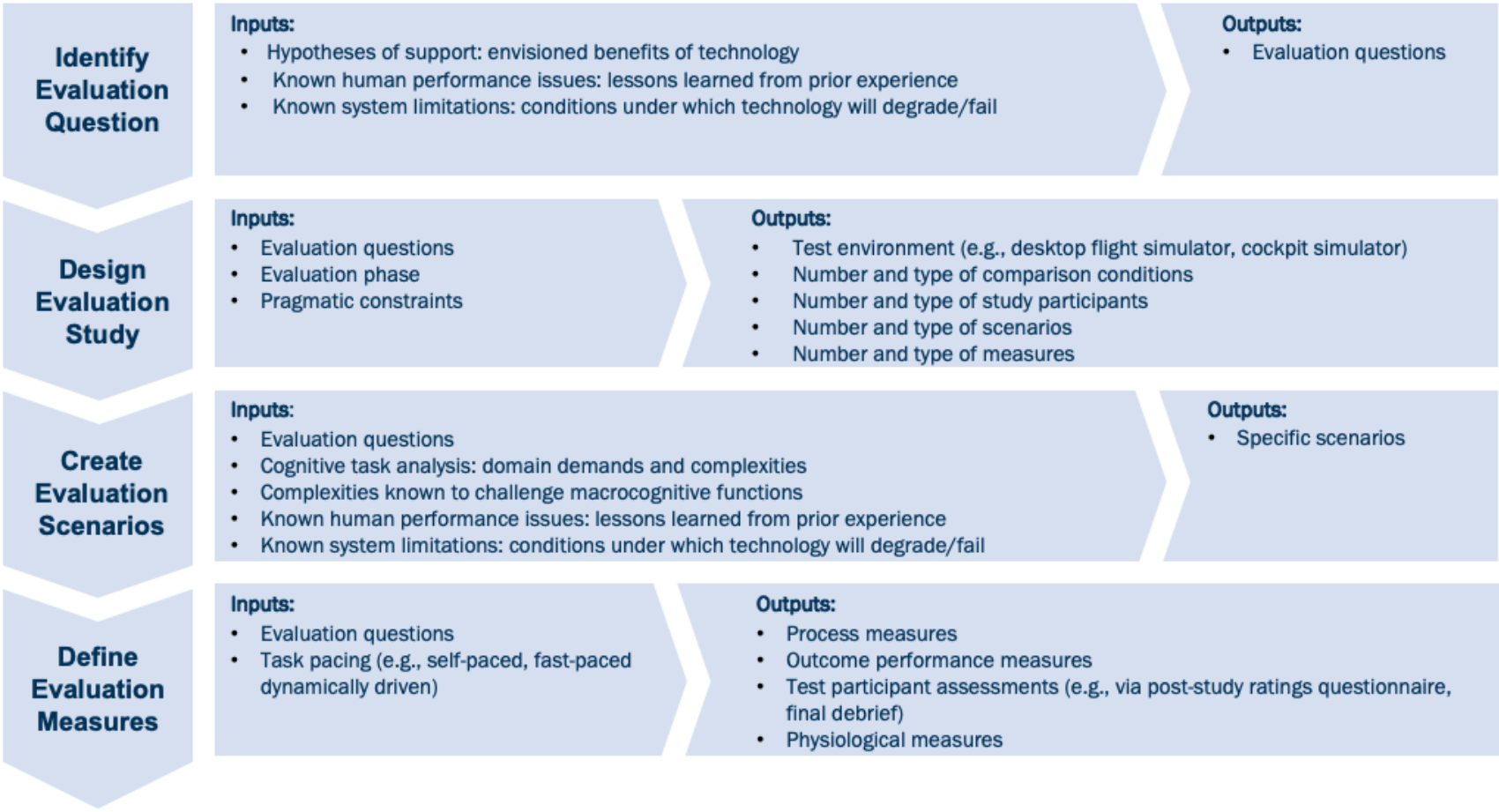

Figure 4 Considerations that go into design of an evaluation study and selection of evaluation measures

- **Human performance issues of concern**: A second consideration relates to concerns with respect to potential negative impacts the new technology may have on pilot macrocognitive functions and performance. Human performance concerns typically come from lessons learned from introduction of prior systems. Concerns may include excessively high workload, ill-timed interruptions that disrupt pilot performance on their primary tasks, or automation bias that causes pilots to follow the recommendations of the aiding technologies even in cases where its recommendations are incorrect. The objective is to ensure that these types of human performance issues that have arisen with similar technologies, do not occur or are mitigated.
- **Known system limitations**: This refers to known or suspected conditions under which the technology is anticipated to fail or perform sub-optimally. The evaluation question of interest is whether the pilot will be able to recognize these 'edge case' situations and appropriately compensate so that the joint person-technology system continues to operate resiliently.

These three categories of considerations are intended as prompts to think about in generating the evaluation questions, there need not necessarily be an evaluation question from each category.

Based on the evaluation questions, maturity of the technology aid being evaluated, the phase of evaluation, and pragmatic constraints such as availability of pilots to serve as participants, an overall ***evaluation study design*** can be specified. Study design would include:

- What test environment will be used for the evaluation. Early in the design development the evaluation might use a rapid prototype of the proposed interface displayed on a desk-top computer in a laboratory setting; as the design matures, a dynamic prototype driven by a flight simulator may be tested using a high-fidelity cockpit simulator. Still later in the design development, the evaluation might take the form of actual helicopter flight tests.
- Whether comparison conditions will be included in the study, and if so, what they will be. Ideally the benefits of a new technology aid will be evaluated by comparing performance with and without the technology aid. Other comparisons might include evaluating the impact on performance of novel display features provided by the new technology independent of the impacts of the new forms of automation and intelligent aiding provided by the new technology. In this case there may be three comparison conditions: display only; automated recommendations only; and display plus automated recommendations.
- The number and type of study participants to be used in the study. Ideally a study evaluating the benefits of a new aiding technology for helicopter pilots would include current military helicopter pilots as the test participants. However, because military helicopter pilots are a limited resource, early evaluations may employ close analog surrogates such as civilian helicopter pilots, or student pilots.
- The number and type of scenarios will be influenced by the evaluation questions. This is described further in the Scenario Design section.
- The number and type of measures. The measures selected will be strongly influenced by the evaluation questions. Ideally the study would include outcome measures, process measures and study participant assessment. This is described further in the Evaluation Questions section.

Scenario Design

A critical element of the study is the scenarios that will be used in the evaluation. Ideally the study would include multiple scenarios that reflect realistic conditions and challenges that are likely to be faced in the actual operational context. Scenarios should include cases that range from straightforward situations that pilots would be able to handle with or without the technology aid; more cognitively challenging cases, where the technology aid is anticipated to provide useful support; and more complex 'edge cases' that are anticipated to challenge the performance of the technology aid. These are included to evaluate the resilience of the joint person-technology system. In particular, the goal is to assess the ability of the pilot to detect that the technology aid has failed or that it is providing a poor recommendation as well as the ability of the pilot to redirect the technology aid toward a more productive solution or to manually take over.

Specific inputs into design of scenarios include:

- Evaluation questions. Scenarios should include conditions relevant to the evaluation questions (e.g., if an evaluation question asks whether the technology aid will be effective under both low and high workload conditions, then there should be high and low workload scenarios).
- Results of cognitive task analyses that reveal domain complexities (e.g., incidents where the events on the ground differed from what was believed based on prior intelligence reports). Often cognitive task analysis elicits a corpus of critical incidents that can be drawn upon to create scenarios.
- Complexities known to challenge the macrocognitive functions. For example, if the technology aid is hypothesized to support sensemaking, then scenarios should be included that challenge such sensemaking (e.g., situations where new information coming in is unexpected or surprising triggering a need for additional sensemaking). Strategies for creating scenarios that challenge different macrocognitive function are discussed in Patterson and colleagues (2010).
- Known human performance issues based on prior experience with similar systems. For example, mode errors are a well-known type of error in aviation systems. To check for this the researchers might include scenarios where there is a change in mode of an automated navigation system to assess whether the pilot is able to detect it.
- Known system limitations (e.g., known situations where the technology is likely to perform less well). For example, known inability to detect certain types of obstacles, or to identify certain types of threats.

Evaluation Measures

The final important element in design of a study is the selection of evaluation measures. Evaluation measures will primarily be dependent on the evaluation questions. Other considerations include task pacing. Some measures, such as think-aloud, may be more appropriate for self-paced tasks vs. event-driven tasks where the participant needs to monitor and rapidly respond to dynamically changing events. Another consideration is whether interruptions and freeze points that are required in the case of query measures of SA such as SAGAT would be acceptable or be disruptive to the ongoing cognitive and collaborative performance. Degree of realism is yet another consideration. For example, if the goal is high realism, it may not be desirable to utilize a secondary task performance measure of workload that requires the pilot to perform an artificial task such as repeating back an arbitrary list of numbers when presented a prompt.

As a general rule, it is desirable to include multiple measures, including measures of objective outcome performance, process measures, and participant assessments. Participant assessments may be obtained via questionnaires presented at the completion of the study and/or final verbal feedback debriefs.

ILLUSTRATIVE EXAMPLE OF A TECHNOLOGY EVALUATION: A LANDING AID FOR THE PILOT FLYING

Below we describe a hypothetical aiding technology, a landing aid for the pilot flying in a FVL context. We then describe how this hypothetical aid could be evaluated organized using the logic outlined in Figure 4 for specifying the design of an evaluation, as well as the evaluation measures to be used.

Description of Hypothetical Aiding Technology

This hypothetical aid is designed to support pilots landing to exfiltrate ground force personnel under fire. This algorithm integrates information on threats, terrain, wind direction/speed, and location information for the soldiers with an existing system that automatically scans the terrain for hazards. The system generates a primary and alternate landing zone. The helmet mounted display depicts the system-generated constraints (e.g., obstructions, hazards, unsafe terrain) and information about the ground force on the pilot's display while the panel mounted display depicts system-generated primary and alternate landing zones as pilot-selectable options on a map overlay. The pilot can:

- Select from the two options for an automated landing to that site,
- Select from the two options and then bias the flight control system for a joint pilot-automation controlled landing (pilot controls certain degrees of freedom while automation controls others), or
- Use the symbology as inputs to an aided landing.

This hypothetical system builds on design concepts identified in Takahashi and colleagues (2021) and Hartnett & Hicks (2021).

Identifying the Evaluation Questions

The evaluation questions are identified based on the hypotheses of support, known human performance issues, and known system limitations. In the case of our hypothetical example these have been identified as follows.

Hypotheses of Support (HOS):

SA and sensemaking: Provides helicopter crew with SA of the overall conditions at the landing location, including status and location of ground force personnel, threats, weather, and potential hazards.

Planning: Generates primary and alternate landing zones as options to the pilot.

***Coordination with Technology*:** It allows the pilots to direct the automated landing aid as well as to take over landing manually, as well as depicting the constraints that were taken into account by the aid.

Coordination with other people: Provides a common picture of the landing site conditions and current plan for landing zone and how it is planned to be approached that can be shared with ground forces as well as with onboard crew chiefs – can look at displays and understand the orientation -- could communicate electronically where it is aimed at landing.

Managing workload: It is anticipated to reduce workload by fusing disparate information into a common picture, offloading the landing zone planning activity, and reducing verbal communication requirements.

Known System Limitations (KSL)

System has limited ability to generate optimal route in the following situations:

- Where threats and wire obstacles are in close proximity,
- If the landing surfaces are unstable (e.g., near water, boggy from recent precipitation),
- If exfiltration force composition changes in the approach phase (meaning that more soldiers must be exfiltrated or different groups of soldiers need exfiltration in the final minutes before the helicopter lands),
- If the threat situation changes.

Human Performance Issues of Concern (HPI)

- Impact on workload.
- Potential for Automation Bias / Impact on Resilience: Can people recognize when the system's recommended landing zone plan is inappropriate and needs to be over-ridden?
- Impact on trust in the technology.
- Impact on coordination with others in the cockpit and on the ground: Can the others understand the landing zone plan and prepare their own actions?

Evaluation Questions

From these considerations the following evaluation questions have been defined as depicted in Table 3. The linkages to HOS, KSL, and HPI are indicated with checkmarks in the respective column.

Table 3 Evaluation questions for the hypothetical landing aid

#	Evaluation Question	HOS	HPI	KSL
1	Does it support improved performance? (better than the person working alone – with respect to identifying appropriate landing zone, appropriately orienting the helicopter for landing, and executing the landing)?	✓		
2	Does it support SA/sensemaking relating to the landing situation and the factors that are important to consider in deciding where and how to land?	✓		
3	Does it enable the person to understand output plans provided by the automation and how they were generated – appropriate mental model? Good explanations?	✓		
4	Does it foster appropriate levels of trust?		✓	
5	Directability – is it easy to modify its plan? Can the pilot designate a different landing site or different landing orientation and still have it provide effective support?	✓	✓	
6	Resilience – ability to operate effectively in both routine situations, challenging situations, and situations beyond the competence of the advisory system		✓	✓
7	Does it enable performance under manageable levels of workload?		✓	
8	Is the information distributed across the helmet mounted display and panel mounted display easy to integrate and understand?	✓	✓	
9	Does it facilitate coordination with others (e.g., crew chief(s) on board and/or commander on the ground coordinating the exfiltration)?	✓	✓	
10	Is it perceived overall to be useful and usable by the study participants? What are opportunities for improvements?	✓	✓	

Proposed Series of Studies to Address the Evaluation Questions

Given the long list of evaluation questions, we determined that these could best be addressed across three studies. Each study would focus on a subset of the evaluation questions and use a progressively more sophisticated testbed. The first study focuses on the ability of the landing aid to support pilot landing decisions. This first study does not consider the specific hardware that will be used to present the information or the fact that landing decisions involve close communication and coordination across a distributed team. Once the value of the landing aid is established, the second study addresses the question of how the information is to be displayed in the cockpit, and more particularly whether distributing information across a panel mounted display and a helmet mounted display is effective. This study is conducted in a full scope helicopter simulation. The last proposed study focuses on the question of whether the landing aid and associated displays facilitates or disrupts teamwork. Specifically, whether the aid facilitates communication and coordination between the pilot flying and other individuals both in cockpit and on the ground as is hypothesized.

Study 1 addresses Questions 1-7 that relate to the effectiveness of the information and planning support provided by the landing aid. It would also cover Question 10 (overall usability and usefulness). These questions can be addressed using a desktop simulator with realistic flight dynamics. The participant would be asked to fly to the site, select an appropriate landing and execute the landing. The performance with the landing aid would be compared to performance using more conventional helicopter cockpit displays. This would allow objective assessment of the impacts (both positive and negative) of the landing aid relative to conventional helicopter cockpit displays.

Study 2 addresses Question 8 that focuses on the specific distribution of information across hardware (i.e., that information on the helmet mounted display needs to be integrated with information on the panel mounted display). It would also cover Question 10 (overall usability and usefulness). This evaluation requires use of more realistic display hardware including a helmet mounted display (for displaying system constraints and casualty information) and a panel mounted display (for displaying system-generated primary and alternate landing zones). It would be most appropriately tested in a helicopter cockpit simulator with realistic hardware displays and controls and flight dynamics but would not require a full motion simulator.

Study 3 addresses Question 9 regarding support provided by the landing aid for coordination among multiple individuals - the pilot flying, the crew chief(s) in the back, and the commander on the ground who is coordinating with the pilot to support landing and exfiltration. It would also cover Question 10 (overall usability and usefulness).

For Study 3 you would need a laboratory environment with multiple linked workstations that allows for distributed multi-person performance. For example, someone could be at one workstation taking on the role of the cockpit pilot-on-the-controls flying toward the landing site; someone else could be at a different workstation taking on the role of a crew chief at the back of the helicopter with their own displays; and someone could be at a third workstation taking on the role of the ground commander. These workstations would be in separate locations that would not allow for direct lines of site or verbal communication. The workstations would need to have displays representative of the proposed landing aid displays intended for each role and would need to have capability for communication among the individuals across the three workstations (e.g., simulated radio communication).

Below we elaborate on Study 1.

Detailed Description of Study 1 Methods and Measures

For illustrative purposes here we elaborate on Study 1 methods and measures. Studies 2 and 3 could be expanded in similar ways.

The purpose of Study 1 is to establish the benefits of the landing aid for the pilot on-the-controls flying. It is designed to address evaluation Questions 1-7, and 10. The set of measures we propose to use for Study 1 and how they relate back to the evaluation questions are summarized in Table 4.

Table 4 Measures selected for each hypothetical landing aid evaluation question

Evaluation Question	Measure(s)
Question 1: Does it support improved performance? (e.g., better than the person working alone – with respect to identifying appropriate landing zone, appropriately orienting the helicopter for landing, and executing the landing)	***Operational performance measures:*** time to generate and accuracy of: • landing zone selected; • orientation of landing selected; • landing execution. Performance would be compared between the conditions where participants are performing manually and the condition where the participant is flying using the landing aid.
Question 2: Does it support SA/sensemaking relating to the landing situation and the factors that are important to consider in deciding where and how to land?	***Measures of SA/Sensemaking:*** • Embedded real-time probes: The pilot would be instructed to indicate to the crew chief (a study confederate) their decision of where and how to land and the factors they considered in making that decision. • Self-report measure of SA collected after each scenario is completed.
Question 3: Does it enable the person to understand output plans provided by the automation and how they were generated – appropriate mental model? Good explanations?	***Measures of Transparency, Understandability, or Explainability:*** • Self-report ratings of explainability at the completion of the study • Measure of mental model of how the landing aid works collected at the completion of the study (e.g.., via self-report questionnaire.)
Question 4: Does it foster appropriate levels of trust?	***Measure of Reliance and Compliance:*** • Computed correct usage (when the landing aid solution is correct) and correct rejection (when the landing aid solution is wrong) • Self-report trust ratings provided by the participants at the completion of the study
Question 5: Directability – is it easy to modify its plan? Can the pilot designate a different landing site or different landing orientation and still have it provide effective support?	***Measures of Directability:*** • Ability of pilot to modify the landing solution generated by the automated landing aid in scenarios where the landing solutions provided by the landing aid would not work • Reaction times/success of the Pilot taking over manual control when the system fails or is beyond its bounds of competence • Self-report ratings of directability of the landing aid provided by the participants at the completion of the study
Question 6: Resilience – ability to operate effectively in both routine situations, challenging situations, and situations beyond the competence of the advisory system	***Operational Performance Measures*** • Comparison of performance in the manual and landing aid conditions on the challenging scenarios where the landing aid solution is not correct or not optimal.
Question 7: Does it enable performance under manageable levels of workload?	***Measures of Workload*** • Physiological measure: heart rate variability • Self-Report Subjective Measure: NASA-TLX completed at the end of each scenario
Question 10: Is it perceived overall to be useful and usable by the study participants? What are opportunities for improvements?	***Measures of Usefulness and Usability*** • Participant assessment ratings of usefulness and usability provided by participants at the completion of the study (usability and usefulness questionnaire) • Final verbal feedback debriefs

Test environment: This study can be conducted using a desktop simulator with realistic flight dynamics. The participant would be asked to fly to the site, select an appropriate landing zone, and execute the landing.

Types of participants: Individuals with helicopter piloting experience who would take on the role of Pilot Flying. We would also include confederates that would take on the role of co-pilot, crew chief(s) and commander on the ground. These individuals would not be participants in the study, they would be part of the experiment team. Their interaction with the pilot would be scripted.

Task to be performed: The test participant serving as the pilot flying would be asked to fly to and land at the location required to pick up the ground force.

Comparison conditions: Two conditions – (1) Aided vs. (2) Manual (as is done today). In the aided condition the pilot would have available to them the landing aid displays including the recommended primary and alternate landing options. In the Manual condition they would have the standard displays and controls available today.

Range of scenarios: The participant would be asked to perform the landing task for multiple scenarios that vary in complexity. The selection of scenarios would be informed by prior cognitive task analyses, hypotheses of support, known system limitations and human performance issues of concern. The goal is to create situations that are representative of realistic complexities that arise in the field, that reflect the types of conditions where the technology aid is expected to provide meaningful support, as well as to stress test the joint person-technology system to determine whether it can continue to function effectively even in cases beyond the capabilities of the technology aid (i.e., to test for resilience).

Scenarios would include:

- Straightforward scenarios– known number and location of soldiers requiring exfiltration, threat that is at a medium range (e.g., threat is not very effective at the landing zone), wires and poles easily distinguishable.
- Straightforward scenarios that can be handled by the landing aid but that are expected to be higher workload for the pilot (e.g., multiple threats, multiple obstacles that need to be avoided)
- Challenging scenarios that reflect known weaknesses or exceed the capabilities of the landing aid. These are situations where the recommended landing options provided by the landing aid would be wrong or suboptimal. The pilot would need to either redirect the landing aid to take advantage of its aid capabilities while specifying where and how to land or take over manual control. The scenario would unfold with injects that invoke known system limitations such as:

    - At two minutes prior to landing to pick up five soldiers, one of the helicopters in the flight has a mechanical malfunction. Therefore, the helicopter must pick up personnel from two groups at the landing zone. The pilot may alter the landing site to support the maneuver of ground force personnel.
    - At two minutes prior to landing after the landing location was selected by the pilot, the landing zone comes under fire from a new direction. Therefore, the joint pilot-system need to take this new threat into account in deciding how and where to land.

- At two minutes prior to landing the pilot is informed that the selected landing surface is unstable, and they need to identify an alternative landing location.
- As the pilot is coming in for landing, the pilot sees that two towers on either side are connected by wires which cross the approach path, and they must identify an alternative landing location.

The suite of measures employed in Study 1 will help answer whether the landing aid improves pilot landing performance across a range of anticipated and unexpected situations under manageable workload, and whether the pilots are able to understand and maintain appropriate levels of trust in the landing aid, knowing when to follow its guidance and when to manually take-over or redirect it. In addition, the usability and usefulness questionnaire and final verbal feedback debrief will provide valuable suggestions for improving the landing aid displays and underlying algorithms. Once the benefits of the landing aid to the pilot on-the-controls who is flying the helicopter is established, Study 2 can be performed to evaluate the proposed distribution of information across the cockpit helmet mounted display and panel mounted display; and Study 3 can be performed to assess the impact of the landing aid on communication and coordination across the distributed team within the cockpit and on the ground that are involved in landing coordination.

8 Summary and Conclusions

The FVL program promises dramatically improved rotorcraft platforms that will support aviators in joint all-domain operations. USAARL is wisely considering the role of the pilot and is working to ensure that technology will aid good decision making and SA on the part of the human operators. An important step in that effort is to create clear guidance about how to evaluate technologies intended to support or even improve decision making and SA. The goal of this report is to support and guide people in the research, development, test, and evaluation disciplines as they create evaluation plans and select methods and measures to better assess the utility and efficacy of potential technologies for the FVL aviator.

The objective of this study was to examine the scientific literature and FVL context to recommend measures and methods to evaluate future technologies that influence pilot decision making and SA. The recommendations described in this report are based on the scientific literature and take into account the FVL context and types of support technologies anticipated for FVL. Our analysis is grounded in a model of decision making described in our earlier report (Roth, Klein & Ernst, 2021) and refined in Section 2 of this report (Figure 1). Built on the macrocognition literature, this model articulates five key areas of cognition that underlie how pilots make decisions. Sensemaking, directing attention, managing workload, planning, and communicating/coordinating are foundational for both rapid intuitive and slower deliberative decision making. This model suggests that these five key areas of cognition should be supported by new tools to be successful.

Looking broadly at the findings from this work, our recommendation is to use scenario-based methods to test and evaluate technologies, making sure to:

- Explore a range of realistic scenarios of use, including cognitively challenging situations and ‘edge cases’ and to,
- Include complementary measures to assess the impact of new technology on workload, SA, and other macrocognitive functions.

To provide concrete examples of our recommendations, we considered two potential technologies and suggested strategies for creating a testing plan, selecting methods, and choosing measures. With the goal of supporting the FVL development process, these examples demonstrated how to: think about a new potential technology, analyze which areas of cognition might be affected, and design research to evaluate that impact.

Looking forward, we offer several suggestions for putting these ideas into practice. First, a theme we noted across many areas of literature is the need for better, more standardized measures tailored to the domain. Applying common measures would help different groups collaborate, enable comparisons over time, and make it easier for designers to set a clear goal. We applaud the many Army researchers who are already working towards this effort. We strongly advocate that the Army continue to codify, operationalize, and validate physiological measures tailored to the FVL context, such as the work of the Operator State Monitoring program (e.g., Vogl et al., 2021). We also recommend initiating efforts to select and tailor measures of human technology interaction

such as measures of trust, explainability, mental models, and interruptability that are important for evaluating technologies in the context of FVL.

We recommend that USAARL conduct an evaluation study of a candidate pilot aiding technology as a means to tailor and exercise the methods and measures described in this report. This evaluation exercise could be conducted by USAARL in collaboration with outside researchers. The process of actually considering and selecting measures, deciding how and what scenarios to create, and implementing an evaluation is an excellent vehicle for establishing best practices. It would provide an opportunity to tailor and codify methods and measures for FVL.

Another recommendation is to leverage the results of this report to develop training workshops and a practitioner handbook to disseminate technology evaluation design best practices. We recommend the Army consider hosting a workshop to discuss best practices for selecting methods and measures, and tailoring the design of an evaluation to specific research goals. An event such as this could help more researchers and technology developers consider the implications of prior research for their own research and evaluation projects. A handbook of best practices for evaluation design, written to support the practitioner, would serve as a complementary strategy for sharing best practices across the FVL community.

In addition, we strongly recommend developing and exercising methods for evaluating integrated systems containing multiple technologies and person-technology interfaces. The application of the Modular Open System approach being used by FVL will allow vendors to independently develop different, potentially competing technologies that must be managed in concert by the pilot. An integrated crewstation will be critical to mitigating cognitive workload. Prior experience with the UH-60 Black Hawk program suggests that developing an integrated crewstation is paramount to supporting the pilot (Ernst et al., 2020). Evaluating individual technologies will not be enough; it will be important to develop and promote a set of evaluation methods to ensure that the integrated system effectively supports the pilots.

References

Alicia, T. J., Hall, B. T., & Terman, M. (2020). Synergistic Unmanned Manned Intelligent Teaming (SUMIT) Final Report. U. S. Army Combat Capabilities Development Command. Technical Report # FCDD-AMT-20-09

Aviation Development Directorate - Eustis. (2018). *Pre-Solicitation Notice for W911W6-19-R-00XX Advanced Teaming Demonstration Program.* U.S. Army Aviation and Missile Research and Engineering Center, Aviation Development Directorate. https://govtribe.com/file/government-file/w911w618r00at-02-attachment-1-advanced-teaming-demonstration-program-description-final-dot-pdf

Bainbridge L. (1983). Ironies of automation. *Automatica*, *19:* 775–779.

Beach, L. R. (1993). Broadening the definition of decision making: The role of prochoice screening of options. *Psychological Science*, 4 (4) 2015 – 2020.

Berry, J. N., Cook, J. L., Ely, C. W., Nelson, C. J., & Riley, P. W. (2021). Hey Larry! Investigating interruptions in Future Vertical Lift Platforms. Naval Postgraduate School Systems Engineering Capstone Report.

Bhaskara, A., Skinner, M., & Loft, S. (2020). Agent transparency: A review of current theory and evidence. *IEEE Transactions on Human-Machine Systems*, 50(3), 215 – 224.

Billman, D., Mumaw, R., & Feary, M.S. (2020). Methods for Evaluating the Effectiveness of Programs to Train Pilot Monitoring. NASA STI Program Report. NASA/TM—20210000045

Billman, D., Zaal, P., Mumaw, R., Lombaerts, T., Torron, I., Jamal, S., Feary, M., (2021, May). Training airline pilots for improved flight path monitoring: The sensemaking model framework. In *Proceedings of the 21st International Symposium on Aviation Psychology*, p 403-408.

Boardman, M. & Butcher, F., (2019). *An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It*. Brussels: North Atlantic Treaty Organization Science and Technology Organization.

Boyd, J. (2018). *A discourse on winning and losing* (pp. 383-385). Maxwell Air Force Base, AL: Air University Press. Retrieved from http://www.airuniversity.af.mil/AUPress/

Bracken, B., Tobyne, S., Winder, A., Shamsi, N., & Endsley, M. (2021). Can situation awareness be measured physiologically? In H. Ayaz, U. Asgher, U., & L. Paletta (Eds) *Advances in Neuroergonomics and Cognitive Engineering.* AHFE 2021. Lecture Notes in Networks and Systems, vol 259. Springer, Cham. https://doi.org/10.1007/978-3-030-80285-1_4

Brand, Y. & Schulte, A. (2021). Workload-adaptive and task-specific support for cockpit crews: Design and evaluation of an adaptive associate system. *Human-Intelligent Systems Integration*, *3*. https://doi.org/10.1007/s42454-020-00018-8

Brunswik, E. (1955). Representative design and probabilistic theory in a functional psychology. *Psychological Review*, *62*: 193 - 217.

Cahour, B., & Forzy, J.-F. (2009). Does projection into use improve trust and exploration? An example with a cruise control system. *Safety Science*, *47*, 1260–1270. https://doi.org/10.1016/j.ssci.2009.03.015

Calhoun, G., Bartik, J., Ruff, H., Behymer, K., & Frost, E. (2021). Enabling human-autonomy teaming with multi-unmanned vehicle control interfaces. *Human-Intelligent Systems Integration*, *3*(2), 155–174. https://doi.org/10.1007/s42454-020-00020-0

Calhoun, G., Ruff, H., Frost, E., Bowman, S., Bartik, J., & Behymer, K. (2021, September). Performance-based adaptive automation: Number of task types and response time measures triggering level of automation changes. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *65*(1): pp. 37-41. Sage CA: Los Angeles, CA: SAGE Publications.

Chiou, E. K., & Lee, J. D. (2021). Trusting automation: Designing for responsivity and resilience. *Human Factors*, 00187208211009995.

Christoffersen, K. & Woods, D. D. (2002). How to make automated systems team players. In E. Salas (Ed.) *Advances in human performance and cognitive engineering research*. UK: Emerald Group Publishing Limited., 1-12

Cooke, N. J., & Gorman, J. C. (2009). Interaction-based measures of cognitive systems. *Journal of Cognitive Engineering and Decision Making, 3*(1), 27-46.

Draper, M., Rowe, A., Douglass, S., Calhoun, G., Spriggs, S., Kingston, D., ... & Frost, E. (2018). *Realizing Autonomy via Intelligent Hybrid Control: Adaptable Autonomy for Achieving UxV RSTA Team Decision Superiority (also known as Intelligent Multi-UxV Planner with Adaptive Collaborative/Control Technologies (IMPACT))*. Air Force Research Laboratory Wright Patterson AFB, OH.

Durso, F. T. & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.) *A Cognitive Approach to Situation Awareness: Theory and Application*. London: Routledge, pp. 137–154.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). *Proceedings of the IEEE 1988 National Aerospace and Electronics Conference.* (3). 789-795

Endsley, M. R. (1995a). Toward a theory of situation awareness in dynamic systems. *Human Factors*, *37*(1), 32-64.

Endsley M. R. (1995b). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65-84.

Endsley, M.R. (2019). Human factors and aviation safety: Testimony to the United States House of Representatives Hearing on Boeing 737-Max8 crashes. Available: https://transportation.house.gov/imo/media/doc/Endsley%20Testimony.pdf.

Endsley M. R. (2020). The divergence of objective and subjective situation awareness: A meta-analysis. *Journal of Cognitive Engineering and Decision Making, 14*(1), 34-53. doi:10.1177/1555343419874248

Endsley, M. R. (2021). *Situation Awareness Measurement: How to Measure Situation Awareness in Individuals and Teams*. Human Factors and Ergonomics Society.

Ernst K., Militello, L., Roth, E., Sushereba, C., & Scheff, S. (2020). *Future Vertical Lift Cognitive Workload Risk Mitigation Study Report* (USAARL-TECH-CR—2021-03). U.S. Army Aeromedical Research Laboratory. https://www.usaarl.army.mil/TechReports/2021-02.pdf

Ernst, K., Militello, L., & Roth, E. (2021). *Future Vertical Lift Complexity in Information Systems Taxonomy of Interactions and Dependencies between Flight and Mission Systems* [unpublished technical report]. Applied Decision Science.

Falkland, E.C., Wiggins, M.W., & Westbrook, J.I. (2020). Cue utilization differentiates performance in the management of interruptions. *Human Factors*, *62*(5). 751-769.

Foroughi, C. K., Devlin, S., Pak, R., Brown, N. L., Sibley, C., & Coyne, J. T. (2021). Near-perfect automation: Investigating performance, trust, and visual attention allocation. *Human Factors*. Advance online publication. https://doi.org/10.1177/00187208211032889

Friesen, D., Borst, C., Pavel, M. D., Masarati, P., & Mulder, M. (2021). Design and Evaluation of a Constraint-Based Helicopter Display to Support Safe Path Planning. In *Nitros Safety Workshop* (pp. 9-11).

Friesen, D., Borst, C., Pavel, M. D., Stroosma, O., Masarati, P., & Mulder, M. (2021). Design and evaluation of a constraint-based head-up display for helicopter obstacle avoidance. *Journal of Aerospace Information Systems*, *18*(3), 80–101. https://doi.org/10.2514/1.I010878

Gawron, V. (2019). *Human Performance, Workload, and Situational Awareness Measures Handbook.* Boca Raton, FL: CRC Press.

Green, D. M. & Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. Oxford: Wiley.

Grundgeiger, T., Michalek, A., Hahn, F., Wurmb, T., Meybohm, P., & Happel, O. (2022). Guiding attention via a cognitive aid during a simulated in-hospital cardiac arrest scenario: A salience effort expectancy value model analysis. *Human Factors*. https://doi.org/10.1177/00187208211060586

Gugerty, L. & Falzetta, M. (2005). Using an event-detection measure to assess drivers' attention and situation awareness. *Proceedings of the 49th Annual Meeting of the Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors & Ergonomics Society.

Hameed, S., Ferris, T., Jayaraman, S., & Sarter, N. (2009). Using informative peripheral visual and tactile cues to support task and interruption management. *Human Factors*, *51*(2). 126-135.

Hartnett, G. & Hicks, J. (2021). *Dynamic InfoGraphics (DIG) for transparency of systems: A concept for transparency of semi-autonomous/fully autonomous (SAFA) and artificial intelligence (AI) systems* [PowerPoint slides]. U.S. Army Combat Capabilities Development Command – Data & Analysis Center (Human Systems Integration Division).

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (Vol. 52, pp. 139–183). North-Holland. https://doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9.

Headquarters, Department of the Army (2019). ADP 5-0: The Operations Process [Army Doctrine Publication].

Hoffman, R. R. & Hancock, P. A. (2017). Measuring resilience. *Human Factors*, *59*(4):564-581. doi:10.1177/0018720816686248

Hoffman, R.R., Johnson, M., Bradshaw, J.M., & Underbrink, A. (2013). Trust in automation. *IEEE Intelligent Systems*, pp. 84-88.

Hoffman, R.R., Klein, G., & Borders. (2018). "A Guide to the Measurement and Evaluation of User Mental Models". Technical Report, DARPA Explainable AI Program.

Hoffman, R.R., Mueller, S.T., Klein, G., & Litman, J. (2018a). "Measuring Trust in the XAI Context." Technical Report, DARPA Explainable AI Program.

Hoffman, R.R., Mueller, S.T., Klein, G., & Litman, J. (2018b). "Metrics for Explainable AI: Challenges and Prospects." Report on Award No. FA8650-17-2-7711, DARPA XAI Program. [https://arxiv.org/abs/1812.04608]

Huang, L., Cooke, N., Johnson, C., Lematta, G., Bhatti, S., Barnes, M., & Holder, E. (2020). *Human-Autonomy Teaming: Interaction Metrics and Models for Next Generation Combat Vehicle Concepts*. Arizona State University, Mesa, AZ

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). Cardiac measures of cognitive workload: A meta-analysis. *Human Factors*, *61*(3), 393-414.

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

Johnson, M., Bradshaw, J. M., & Feltovich, P. J. (2018). Tomorrow's human–machine design tools: From levels of automation to interdependencies. *Journal of Cognitive Engineering and Decision Making*, *12*(1), 77–82. https://doi.org/10.1177/1555343417736462

Kahneman, D. (2011). *Thinking fast and slow*. New York, NY: Farrar, Straus, & Giroux.

Kahneman, D. & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica, 47* (2), 263-292.

Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2021). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 00187208211013988.

Kinney, L. & O'Hare, D. (2020). Responding to an unexpected in-flight event: Physiological arousal, information processing, and performance. *Human Factors*, *65*(2), 737-750.

Klein, D. E., Klein, H. A., & Klein, G. (2000). Macrocognition: Linking cognitive psychology and cognitive ergonomics. In *Proceedings of the 5th international conference on human interactions with complex systems* (pp. 173-177).

Klein, G. (1989). Recognition-primed decisions. In W. B. Rouse (Ed.) *Advances in Man-Machine Systems Research, vol. 5*, Greenwich, CT: JAI Press, 47-92.

Klein, G., Borders, J., Hoffman, R.R. & Mueller, S.T. (2021). "A Method for Evaluating Users' Understanding of XAI Systems: The Mental Model Matrix. Technical Report, DARPA Explainable AI Program.

Klein, G., Feltovich, P. J., Bradshaw, J. M., & Woods, D. D. (2005). Common ground and coordination in joint activity. In *Organizational Simulation*. https://doi.org/10.1002/0471739448.ch6

Klein, G., Hoffman, R.R., & Mueller, S.T. (2021). "Scorecard for Self-Explaining Capabilities of AI Systems" Technical Report, DARPA Explainable AI Program.

Klein, G., Jalaeian, M., Hoffman, R.R., Mueller, S.T., & Clancey, W.J. (2021). "Requirements for the Empirical Assessment of Human-AI Work Systems: A Contribution to AI Measurement Science." Technical Report, DARPA Explainable AI Program.

Klein, G., Ross, K. G., Moon, B. M., Klein, D. E., Hoffman, R. R., & Hollnagel, E. (2003). Macrocognition. *IEEE Intelligent Systems*, *18*(3), 81–85. https://doi.org/10.1109/MIS.2003.1200735

Klein, G., Woods, D. D., Bradshaw, J. M., Hoffman, R. R., & Feltovich, P. J. (2004). Ten challenges for making automation a "team player" in joint human-agent activity. *IEEE Intelligent Systems*. https://doi.org/10.1109/MIS.2004.74

Landman, A., Groen, E.L., van Paassen, M.M., Bronhurst, A.W., & Mulder, M. (2017), Dealing with unexpected events on the flight deck: A conceptual model of startle and surprise. *Human Factors*, *59*(8), 1161-1172.

Landman, A., van Oorschot, P., van Paassen, M. M., Groen, E., Bronkhorst, A., & Mulder, M. (2018). Training pilots for unexpected events: A simulator study on the advantage of unpredictable and variable scenarios. *Human Factors*, *60*(6), 793-805.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, *46*(1), 50-80.

Loft, S., Sadler, A., Braithwaite, J., & Huf, S. (2015). The Chronic Detrimental Impact of Interruptions in a Simulated Submarine Track Management Task. *Human Factors*, 57(8). 1417-1426.

Lu, S.A., Wickens, C.D., Prinet, J.C., Hutchins, S.D., Sarter, N., & Sebok, A. (2013). Supporting interruption management and multimodal interface design: Three meta-analyses of task performance as a function of interrupting task modality. *Human Factors, 55*(4), 697-724.

Lyons, J. B., Stokes, C. K., Eschleman, K. J., Alarcon, G. M., & Barelka, A. J. (2011). Trustworthiness and IT suspicion: An evaluation of the nomological network. *Human Factors*, *53*(3), 219-229.

Martin, P., Calhoun, P., & Schnell, T. (2019). *Objective measures of pilot workload: Study results* (Contract No. 18S0234C). 412th Test Wing, Edwards Air Force Base, CA: Air Force Materiel Command.

Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, *85*(2), 273.

Matthews, G. & Reinerman-Jones, L. E. (2017). *Workload Assessment: How to diagnose workload issues and enhance performance. Users' Guides to Human Factors and Ergonomics Methods*, Santa Monica, CA: The Human Factors and Ergonomics Society.

McNeese, N. J., Demir, M., Cooke, N. J., & She, M. (2021). Team situation awareness and conflict: A study of human–machine teaming. *Journal of Cognitive Engineering and Decision Making*, *15*(2-3), 83-96.

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. (2011). Affective processes in human-automation interactions. *Human Factors*, *53*(4), 356-370.

Metzger, U. & Parasuraman, R. (2005). Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors*, *47*(1), 35–49.

Meyer, J. (2004). Conceptual issues in the study of dynamic hazard warnings. *Human Factors*, *46*(2), 196-204.

Militello, L. G., Roth, E. M., Scheff, S., Ernst, K., Sushereba, C., Diiulio, J., & Klein, D. (2019a). Toward an optimally crewed future vertical lift vehicle: Crewing strategy and recommendations. Applied Decision Science, LLC. [Unpublished technical report, Contract No. NNX16AJ91A, 21-1614-5637-ADS2018].

Militello, L. G., Roth, E. M., Scheff, S., Ernst, K., Sushereba, C., Klein, D., & Wonderly, S. (2019b). Crew configuration analysis for future airborne reconnaissance operations (unpublished technical report, Contract No. NNX16AJ91A, 21-1614- 5637- ADS2018). NASA AMES.

Militello, L.G., Roth, E.M., Scheff, S., Ernst, K.M., Sushereba, C.E., Diiulio, J.B., & Klein, D. (2018). Toward an optimally crewed future vertical lift vehicle: Overview of the envisioned world, core missions, and pertinent

technology. Applied Decision Science, LLC. [Unpublished technical report, Contract No. NNX16AJ91A, 21-1614- 5637-ADS2018].

Miller, J. D., Godfroy-Cooper, M., & Szoboszlay, Z. (2021). Degraded visual environment mitigation (DVE-M) program: Bumper Radar obstacle cueing flight trials 2020. Presented at the *Vertical Society's 77th Annual Forum & Technical Display*, Virtual Conference, May 10-14, 2021.

Moacdieh, N.M., Devlin, S.P., Jundi, H., & Riggs, S.L. (2020). Effects of workload and workload transitions on attention allocation in a dual-task environment: Evidence from eye tracking metrics. *Journal of Cognitive Engineering and Decision Making*. 14(2). 132-151.

Mosier, K.L., Sethi, N., McCauley, S., Khoo, L., & Orasanu, J. (2007). What you don't know can hurt you: Factors impacting diagnosis in the automated cockpit. *Human Factors, 49*(2), 300-310.

Mumaw, R., Billman, D., & Feary, M.S. (2020). Analysis of Pilot Monitoring Skills and a Review of Training Effectiveness. NASA STI Program Report. NASA/TM–20210000047.

National Academies of Sciences, Engineering, and Medicine (2021). *Human-AI Teaming: State of the Art and Research Needs*. Washington, DC: The National Academies Press. https://doi.org/10.17226/26355

National Research Council (NRC) (2007). *Human-System Integration in the System Development Process: A New Look.* National Academy Press.

Neville, K., Rosso, H., & Pires, B. (2021). A systems-resilience approach to technology transition in high-consequence work systems. *Proceedings of the Naturalistic Decision Making and Resilience Engineering Symposium*, Toulouse, France.

Norman, D. A. (1990). The "Problem" with automation: Inappropriate feedback and interaction, not "Over-Automation." *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *327*(1241), 585–593. http://www.jstor.org/stable/55330

Olmos, O., Wickens, C. D., & Chudy, A. (2000). Tactical displays for combat awareness: An examination of dimensionality and frame of reference concepts and the application of cognitive engineering. *The International Journal of Aviation Psychology*, *10*(3), 247–271. https://doi.org/10.1207/S15327108IJAP1003_03

Ophir-Arbelle, R., Oron-Gilad, T., Borowsky, A., & Parmet, Y. (2013). Is more information better? How dismounted soldiers use video feed from unmanned vehicles: Attention allocation and information extraction considerations. *Journal of Cognitive Engineering and Decision Making*, *7*(1), 26-48.

Patterson, E.S., Roth, E.M., & Woods, D.D. (2010). Facets of complexity in situated work. in J. E. Miller & E. S. Patterson (Eds.) *Macrocognition Metrics and Scenarios: Design and Evaluation for Real World Teams*, Boca Raton, FL: CRC Press, pp. 221 - 252.

Rankin, A., Woltjer, R., & Field, J. (2016). Sensemaking following surprise in the cockpit—A re-framing problem. *Cognitive Technology & Work, 18*: 623-642.

Rasmussen, J. L. (1976). Outlines of a hybrid model of the process plant operator. In T. B. Sheridan & G. Johannsen (Eds) *Monitoring Behavior and Supervisory Control* (pp. 371-384). New York: Plenum Press.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive Systems Engineering.* New York: Wiley and Sons.

Ratwani, R. & Trafton, J.G. (2010). An eye movement analysis of the effect of interruption modality on primary task resumption. *Human Factors, 52*(3), 370-380.

Roscoe, A. H. A. & Ellis, G. A. (1990). A subjective rating scale for assessing pilot workload in flight: A decade of practical use. *Royal Aircraft Establishment TR 90019*.

Roth, E. M., Bennett, K., & Woods, D. D. (1987). Human interaction with an 'intelligent' machine. *International Journal of Man-Machine Studies*, *27*, 479-525.

Roth, E. M., Bisantz, A. M., Wang, X., Kim, T., & Hettinger, A. Z. (2021). A work-centered approach to system user-evaluation. *Journal of Cognitive Engineering and Decision Making*, 15(4), 155–174. https://doi.org/10.1177/15553434211028474

Roth, E. M., DePass, B., Scott, R., Truxler, R., Smith, S., & Wampler, J. (2017). Designing collaborative planning systems: Putting joint cognitive systems principles to practice. In P. J. Smith & R. R. Hoffman (Eds). *Cognitive Systems Engineering: The Future for a Changing World*. Boca Raton: Taylor & Francis, CRC Press. (247-268).

Roth, E., Klein, D., & Ernst, K. (2021). *Aviation Decision Making and Situation Awareness Study: Decision Making Literature Review* (USAARL-TECH-CR—2022-17). U.S. Army Aeromedical Research Laboratory. https://www.usaarl.army.mil/assets/docs/techReports/2022-17.pdf

Roth, E.M. & Eggleston, R.G. (2018). Forging new evaluation paradigms: Beyond statistical generalization. In J. E. Miller & E. S. Patterson (Eds) *Macrocognition Metrics and Scenarios: Design and Evaluation for Real World Teams.* Boca Raton, FL: CRC Press, pp. 203 - 220.

Roth, E. M., Sushereba, C., Militello, L. G., Diiulio, J., & Ernst, K. (2019). Function allocation considerations in the era of human autonomy teaming. *Journal of Cognitive Engineering and Decision Making*, *13*(4), 199–220.

Sarter, N.B. & Woods, D.D. (1995). How in the world did I ever get into that mode: Mode error and awareness in supervisory control. *Human Factors*, 37(1), 5–19.

Schaefer, K. E. (2013). The perception and measurement of human-robot trust. Doctoral dissertation, University of Central Florida, Orlando, FL.

Schnell, T., Keller, M., & Poolman, P. (2008). Neurophysiological workload assessment in flight. In *2008 IEEE/AIAA 27th Digital Avionics Systems Conference* (pp. 4.B.2-1 to 4.B.2-14). IEEE.

Schriver, A.T., Morrow, D.G., Wickens, C.D., & Talleur, D.A. (2008). Expertise differences in attentional strategies related to pilot decision making. *Human Factors*, 50(6). 864-878.

Sebok, A. & Wickens, C. D. (2017). Implementing lumberjacks and black swans into model-based tools to support human-automation interaction. *Human Factors*, *59*(2), 189–203. https://doi.org/10.1177/0018720816665201

Selcon, S. J. & Taylor, R. M. (1990). Evaluation of the situational awareness rating technique (SART) as a tool for aircrew systems design. In *Situational awareness in aerospace operations* (AGARD-CP-478) (pp. 5/1–5/8). Neuilly Sur Seine, France: North Atlantic Treaty Organization-Advisory Group for Aerospace Research and Development.

Semmens, R., Martelaro, N., Kaveti, P., Stent, S., & JU, W. (2019). Is now a good time? An empirical study of vehicle-driver communication timing. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019)*, May 4– 9, 2019, Glasgow, Scotland UK. ACM, New York, NY, USA, 12 pages.

Shattuck, L. & Miller, N. (2006). Extending naturalistic decision making to complex organizations: A dynamic model of situated cognition. *Organization Studies*, *27*(7). 989-1009.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 27(3), 360-371.

Smith, P. J., McCoy, C. E., & Layton, C. (1997). Brittleness in the design of cooperative problem-solving systems: The effects on user performance. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, *27*(3), 360–371, 360-371. https://doi.org/10.1109/3468.568744

Stowers, K., Kasdaglis, N., Rupp, M. A., Newton, O. B., Chen, J. Y. C., & Barnes, M. J. (2020). The IMPACT of agent transparency on human performance. *IEEE Transactions on Human-Machine Systems, 50*(3), 245–253. https://doi.org/10.1109/THMS.2020.2978041

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication.* Cambridge University Press.

Suh, Y. & Ferris, T.K. (2019). On-road evaluation of in-vehicle interface characteristics and their effects on performance of visual detection on the road and manual entry. *Human Factors*, *61*(1), 105-118.

Szoboszlay, Z., Miller, J., Godfroy-Cooper, M. (2021). The design of pilot cueing for the degraded visual environment mitigation (DVE-M) system for rotorcraft. Presented at the *Vertical Flight Society's 77th Annual Forum & Technology Display*, virtual, May 11

Takahashi, M. D., Fujizawa, B. T., Lusardi, J. A., Whalley, M. S., Goerzen, C. L., Schulein, G. J., … Waldman, D. W. (2021). Autonomous guidance and flight control on a partial-authority Black Hawk helicopter. *Journal of Aerospace Information Systems*, *18*(10), 686–701. https://doi.org/10.2514/1.I010880

Takahashi, M. D., Goerzen, C. L., Whalley, M. S., Mansur, M. H., Schulein, G. J., Minor, M. A. J., ... & Morford, M. A. J. (2018). Full-scale flight-test results for a rotorcraft safe landing area determination algorithm for autonomous and piloted landing approaches. *Journal of the American Helicopter Society*, 63(4), 1-15.

Takahashi, M. D., Whalley, M. S., Mansur, H., Ott, L. C., Minor, M. J. S., & Morford, M. Z. G. (2016). Autonomous rotorcraft flight control with multilevel pilot interaction in hover and forward flight. *Journal of the American Helicopter Society*, 62(3), 1-13.

Tenney, Y. J. & Pew, R. W. (2006). Situation awareness catches on: What? So What? Now What? *Reviews of Human Factors and Ergonomics*. 2(1):1-34. doi:10.1177/1557234X0600200102

Trapsilawati, F., Chen, C. H., Wickens, C. D., & Qu, X. (2021). Integration of conflict resolution automation and vertical situation display for on-ground air traffic control operations. *The Journal of Navigation*, *74*(3), 619-632.

Trapsilawati, F., Herliansyah, M. K., Nugraheni, A. S., Fatikasari, M. P., & Tissamodie, G. (2020). EEG-based analysis of air traffic conflict: Investigating controllers' situation awareness, stress level and brain activity during conflict resolution. *The Journal of Navigation*, *73*(3), 678-696.

Trapsilawati, F., Wickens, C., Chen C-H, & Qu, X. (2017). Transparency and conflict resolution automation reliability in Air Traffic Control. *19th International Symposium on Aviation Psychology*, Dayton, Ohio, pp. 419-424.

Trapsilawati, F., Wickens, C. D., Qu, X., & Chen, C. H. (2016). Benefits of imperfect conflict resolution advisory aids for future air traffic control. *Human Factors*, *58*(7), 1007-1019.

Vicente, K., Mumaw, R. & Roth, E. (2004). Operator monitoring in a complex dynamic work environment: a qualitative cognitive model based on field observations. *Theoretical Issues in Ergonomics Science*, 5(5), 359-384.

Vogelpohl, T., Gehlmann, F., & Vollrath, M. (2020). Task interruption and control recovery strategies after take-over requests emphasize need for measures of situation awareness. *Human Factors*, *62*(7), 1190-1211.

Vogl, J., Delgado-Howard, C., McAtee, A., Aura, C., Hayes, A., & St. Onge, P. (2021). *Annotated bibliography: Cognitive workload of aviators with various levels of experience.* (USAARL-TECH-BB-2022-13). United States Army Aeromedical Research Laboratory.

Whalley, M. S., Takahashi, M. D., Fletcher, J. W., Moralez III, E., Ott, L. C. R., Olmstead, L. M. G., ... & Conrad, B. (2014). Autonomous Black Hawk in flight: Obstacle field navigation and landing-site selection on the RASCAL JUH-60A. *Journal of Field Robotics*, 31(4), 591-616.

Wiggins, S. L. & Cox, D. A., (2010). System evaluation using the cognitive performance indicators. In J. E. Miller & E. S. Patterson (Eds.) *Macrocognition Metrics and Scenarios: Design and Evaluation for Real World Teams.* Boca Raton, FL: CRC Press, pp. 285 - 302.

Woods, D. (2015). Four concepts for resilience and the implications for the future of resilience engineering. *Reliability Engineering & System Safety*, *141*. https://doi.org/10.1016/j.ress.2015.03.018

Woods, D. D. & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering.* Boca Raton, FL: CRC Press.

Zhang, T., Yang, J., Liang, N., Pitts, B. J., Prakah-Asante, K. O., Curry, R., ... & Yu, D. (2020). Physiological measurements of situation awareness: A systematic review. *Human Factors*, https://doi.org/10.1177/0018720820969071

Appendix

SECOND ILLUSTRATIVE EXAMPLE: INTEGRATED ALE MANAGEMENT SYSTEM

System Description:

This hypothetical aid is an integrated ALE management system supporting reconnaissance pilots. Specifically, this aid has been developed to support a pilot commanding ALEs that is performing reconnaissance in support of a ground maneuver element. The system supports the pilot by ingesting data received from the ALEs and the common operating picture (COP) and presenting a consolidated picture of the area of interest. The system elicits the pilot's goals and priorities to task, monitor, and redirect ALEs. Pilots use panel-mounted and helmet-mounted displays to visualize the information, receive auditory alerts, and can interact with the system via touch (panel mounted display, buttons on the flight control inceptors).

This system builds on ALE management technology designs including IMPACT (Draper et al., 2018), SUMIT (Alicia et al., 2020) and the Advanced Teaming Demonstration Program (Aviation Development Directorate – Eustis, 2018)

Hypotheses of Support (HOS):

SA and Sensemaking: Presenting the data via a map on the panel-mounted display and the helmet mounted display supports pilot sensemaking. Providing selectable presentation of data not only from the ALEs but also other information sources, allows the pilot a more thorough understanding of the battlespace.

Planning: The system elicits the pilot's goals and priorities to automatically plan and replan ALE tasking. The automation provides alternative plans for allocating the ALEs in response to dynamically changing conditions. The pilot can accept alternative, modify, or take over manual control.

Managing attention: The system fuses information from the various inputs and uses multi-modal cueing to make important or time sensitive elements more salient to the pilot. The system alerts the pilot when his/her attention is needed to be aware of an important change, approve an automated decision, or to intervene.

Managing workload: The system facilitates human intervention and control along a continuum, where pilots are able to easily take command of individual ALE payloads and supervise the tasking and execution of many ALEs.

Coordinating with Technology: The system presents information to the pilot on a panel-mounted display and helmet mounted display. Map-based interfaces display the progress and status of ALE

task execution and reconnaissance findings. Pilots can draw on the map display to interact with the ALE Management System. Pilots can accept alternative system generated plans for allocating ALEs, modify them, or take over manual control of ALEs.

Inputs to Evaluation Questions:

Known System Limitations (KSL):

- The system degrades when dynamic objects of interest are dispersed over distances of greater than 5km and will therefore not always generate an optimal plan.
- The system will not be aware of new areas of interest for reconnaissance that communicated to pilot over radio.
- The system will not be aware of ALE malfunctions that may limit their reconnaissance capabilities.

Human Performance Issues of Concern (HPI):

- Multiple alerts and system interruptions may degrade pilot performance on own reconnaissance tasks.
- Potential for Automation Bias / Impact on Resilience: Can people recognize when the system's recommended landing zone plan is inappropriate and needs to be over-ridden
- Impact on workload.
- Impact on trust in the technology.

Evaluation Questions

Table 5 Evaluation questions for ALE management system hypothetical aid

#	Evaluation Question	HOS	HPI	KSL
1	Does it support improved performance? (better than the person working alone – with respect to task, monitor, and redirect ALEs in support of reconnaissance)?	✓		
2	Does it support SA/sensemaking relating to location, status and tasking of ALEs?	✓		
3	Does it enable the person to understand output plans provided by the automation and how they were generated – appropriate mental model? Good explanations?	✓		
4	Does it foster appropriate levels of trust?		✓	
5	Directability – is it easy to modify its plan? Can the pilot designate alternative ALE tasking based on their own assessment?		✓	
6	Resilience – Is the joint person-technology system able to operate effectively in both routine situations, challenging situations, and situations beyond the competence of the advisory system?		✓	✓
7	What is the impact of interruptions by the ALE Management System on the ability of the pilot commanding ALEs to perform their own reconnaissance and communication responsibilities?		✓	

8	Does it enable performance under manageable levels of workload?		✓	
9	Is it perceived overall to be useful and usable by the study participants? What are opportunities for improvements?	✓		

Proposed Study/Studies to Address these Evaluation Questions

Study 1: Addresses questions 1- 6 that relate to the effectiveness of the information and planning support provided by the ALE Management System to the pilot commanding the ALEs. It would also address Question 9 (overall usability and usefulness). These questions do not require a flight simulator since the main questions relate to the ability to task, monitor, and redirect ALEs and does not directly involve flying a helicopter. The study could be conducted using a workstation with displays that can depict the location of the ALEs and their status, the results of their reconnaissance, and the displays required to interact with the ALE Management System.

Study 2: This study focuses more specifically on the impact of interacting with the ALE management system on the ability of the pilot commanding the ALEs to perform their own reconnaissance and communication responsibilities. It would specifically address the impact of interruptions by the ALE Management System on the performance of the Pilot Commanding the ALEs on their own tasks (Question 7) as well as the impact of the additional workload imposed by the ALE management system on the overall pilot workload and ability of the Pilot Commanding the ALE to perform their own tasks (Question 8). It would also address Question 9 (overall usability and usefulness).

Questions 7 & 8 do not necessarily require use of a flight simulator. However, since the pilot commanding the ALE would presumably need to interact with the pilot flying in order to support their own reconnaissance task, it would be ideal if the study were conducted in a cockpit simulator, with a study confederate serving as the pilot flying. The pilot flying would not be a subject of the study. They would fly in a scripted manner, communicating with and following the directions of the pilot commanding the ALE. The pilot commanding the ALEs would be asked to perform multiple tasks simultaneously: (1) perform their own reconnaissance task; (2) communicate with the pilot flying in support of the reconnaissance task; (3) communicate with Commanders on the ground in support of the reconnaissance task; (3) manage the ALEs in support of the reconnaissance task (with the aid of the ALE Management System).

Below we elaborate on both Study 1 and Study 2

Study 1 (Evaluation Questions 1-6, 8 and 9)

Study 1 is intended to answer research questions relating to whether the ALE Management System is providing effective support in tasking and redirecting ALEs. It does not address issues relating to its impact on other ongoing tasks that the Pilot in Command of the ALE is responsible for. These are addressed in Study 2.

Test Environment: This study can be conducted using a desktop computer that can display the reconnaissance task being supported by the ALE and the outputs of the ALE Management System.

Types of Participants: Individuals with helicopter piloting experience who would take on the role of Pilot Commanding the ALEs.

Tasks to be Performed: Task ALEs in support of a reconnaissance task. Tasking would occur both at the start of a reconnaissance mission, and during the reconnaissance mission in response to changing conditions and priorities.

Comparison condition(s): Two conditions – (1) Aided vs. (2) Manual. In the aided condition the pilot would have the ALE Management System to support tasking the ALEs. In the manual condition they would have displays depicting the reconnaissance situation and the location and status of the ALEs but would have to decide how to task the ALEs on their own without the planning support capabilities of the ALE Management System.

Range of Scenarios: The participant would be asked to perform the ALE allocation task for multiple scenarios that vary in complexity. These would include:

- Straightforward scenarios that the ALE Management System is able to allocate appropriately.
- More complex scenarios that require more complicated planning that the ALE management System is able to allocate appropriately (e.g., the best ALE to assign to a reconnaissance task is not the closest one).
- Challenging Scenarios that reflect known weaknesses or exceed the capabilities of the ALE Management System so that the Pilot needs to recognize that they need to modify the solution or take over manually:

    - Dynamic objects of interest are dispersed over distances of greater than 5km and will therefore not always generate an optimal plan.
    - New areas of interest for reconnaissance that are communicated to pilot over radio.
    - ALE malfunctions that limit their reconnaissance capabilities that the ALE Management System does not know about.

Table 6 Measures Selected for Each Evaluation Question-Study 1 Evaluating Hypothetical ALE management aid

Evaluation Question	Measure(s)
Question 1: Does it support improved performance? (better than the person working alone – with respect tasking and redirecting ALEs in support of reconnaissance)?	**Operational performance measures:** Time to generate and accuracy of allocation of ALEs to reconnaissance tasks. Performance would be compared between the conditions where participants are allocating ALEs manually and the condition where the participant is using the ALE management system
Question 2: Does it support SA/sensemaking relating to location, status, and tasking of ALEs	**Measures of SA/Sensemaking:** • SPAM queries (since this is a self-paced task with no confederate acting as flying pilot, the embedded real time probe is less appropriate) • Self-report measure of SA collected at the completion of the study.
Question 3: Does it enable the person to understand output plans provided by the automation and how they were generated – appropriate mental model? Good explanations?	**Measures of Transparency, Understandability, or Explainability:** • Self-report ratings of explainability at the completion of the study • Measure of mental model of how the ALE Management System works collected at the completion of the study, via a Mental Model Matrix exercise.
Question 4: Does it foster appropriate levels of trust?	**Measure of Reliance and Compliance:** • Computed correct usage (when the ALE Management System solution is correct) and correct rejection (when the ALE management system is wrong) • Self-report trust ratings provided by the participants at the completion of the study
Question 5: Directability – is it easy to modify its plan? Can the pilot designate alternative ALE tasking based on their own assessment?	**Measures of directability:** • Ability of pilot to modify the solution generated by the ALE Management System in scenarios where its solution would be suboptimal • Reaction times/success of the Pilot taking over manual control when the system fails or is beyond its bounds of competence • Self-report ratings of directability of the landing aid provided by the participants at the completion of the study
Question 6: Resilience – ability to operate effectively in both routine situations, challenging situations, and situations beyond the competence of the advisory system?	**Operational Performance Measures:** • Comparison of performance in the manual and ALE Management System aiding conditions on the challenging scenarios where the ALE Management system solution is not correct or not optimal
Question 8: Does it enable performance under manageable levels of workload?	**Measures of Workload** • Self-Report Subjective Measure: NASA-TLX completed at the end of each scenario
Question 9: Is it perceived overall to be useful and usable by the study participants? What are opportunities for improvements?	**Measure of Usefulness and Usability** • User assessment ratings of usefulness and usability provided by participants at the completion of the study. • Final verbal feedback debriefs

Study 2 (Evaluation Questions 7-9)

Study 2 addresses the question of how the additional demands imposed by the ALE Management System impact the ability of the Pilot Commanding the ALE to perform their own reconnaissance task and associated communication duties. Specifically, it examines the impact of needing to interact with the ALE Management System and monitor multiple ALE on workload as well as the impact of interruptions caused by requests from the ALE Management System on ongoing reconnaissance performance.

Since the Pilot Commanding the ALE would not be able to monitor and manage multiple ALE manually as well as perform their own reconnaissance task, we do not include a 'manual' condition.

Test environment: Since the Pilot Commanding the ALE needs to interact with the pilot flying in order to support their own reconnaissance task, the study should be conducted in a cockpit simulator, with a study confederate serving as the Pilot flying. The confederate serving as the pilot flying would fly in a scripted manner, communicating with and following the directions of the Pilot Commanding the ALE.

Types of participants: Individuals with helicopter piloting experience who would take on the role of Pilot Commanding the ALEs.

Tasks to be performed: The Pilot Commanding the ALEs would be asked to perform multiple tasks simultaneously: (1) perform their own reconnaissance task; (2) communicate with the pilot flying in support of the reconnaissance task; (3) communicate with Commanders on the ground in support of the reconnaissance task; In the Aided condition they would also be asked to manage the ALEs in support of the reconnaissance task using the ALE Management System.

Comparison condition(s): Two conditions – (1) Aided Management of ALE plus own reconnaissance task versus (2) Control condition (own reconnaissance task only).

In the aided management of ALE task, the pilot would have the ALE Management System to support tasking the ALEs. In addition, they would be performing their own reconnaissance task including communicating with the pilot flying and the Commander on the ground. In the control condition they would only be performing their own reconnaissance task, including communicating, but they would not be managing the ALEs.

Range of Scenarios (guided by prior cognitive task analyses; hypotheses of support; known system limitations; and human performance issues of concern): Same as Study 1 plus consider adding a task requiring the Pilot to move attention to an urgent task outside the ALE Management System. For example, someone outside the helicopter communicates with the Pilot with some urgency, requiring verbal communication back. Time to change focus and respond would be measured.

Table 7 Measures Selected for Each Evaluation Question- Study 2 Evaluating Hypothetical ALE management aid

Evaluation Question	Measure(s)
Question 7: What is the impact of managing ALEs with the support of the ALE Management System on the ability to perform their own reconnaissance task?	**Operational performance measures:** • Time to and accuracy of detecting targets as part of their own reconnaissance task. • Performance would be compared between the conditions where participants are performing
Question 7: What is the impact of the ALE management System on Pilot SA and Sensemaking relating to their own reconnaissance tasks?	**Measures of SA/Sensemaking:** • SAGAT (because it doesn't add to workload in an already high workload situation as compared to SPAM or embedded real-time probes) • Self-report measure of SA collected at the completion of the study.
Question 7: What is the impact of interruptions by the ALE Management System on the ability of the pilot commanding ALEs to perform their own reconnaissance and communication responsibilities? What is the impact of interruptions by other team members or systems on the ability of the pilot to command the ALEs?	**Measures of Impact of Interruptions:** • Lag times to transition and resume tasks following an interruption by the ALE Management System • Lag times to transition and resume tasks following an interruption by another task outside the ALE Management System (i.e., another team member)
Question 8: Does it enable performance under manageable levels of workload?	**Measures of Workload** • Physiological measure: heart rate variability • Self-Report Subjective Measure: NASA-TLX completed at the end of each scenario
Question 9: Is it perceived overall to be useful and usable by the study participants? What are opportunities for improvements?	**Measure of Usefulness and Usability** • User assessment ratings of usefulness and usability provided by participants at the completion of the study • Final verbal feedback debriefs

About the Authors

Emilie M. Roth is owner and principal scientist of Roth Cognitive Engineering. Dr. Roth is a cognitive psychologist by training and has over 30 years of experience in cognitive analysis and design across a broad range of domains including military command control, intelligence analysis, healthcare, and transportation. She has supported design of first-of-a-kind systems including next-generation nuclear power plant control rooms; and work-centered support systems for airlift planning and monitoring for USTRANSCOM and the Air Mobility Command. She is a fellow of the Human Factors and Ergonomics Society and currently serves as a member of the Board on Human-Systems Integration at the National Academies.

Devorah Klein is the owner and principal researcher at Marimo Consulting, LLC. She has a background in cognitive psychology and has worked at the intersection of research and design for over two decades, in highly applied complex settings such as healthcare and aviation. While at IDEO she pioneered work on design for medication adherence that was grounded in cognitive psychology. She received her Ph.D. from the University of Illinois at Urbana-Champaign.

Christen E. Sushereba is a research associate at Applied Decision Science, LLC. For the past ten years, Ms. Sushereba has applied cognitive engineering and human factors methods to a variety of domains including military pararescue, emergency response, cyber security, electronic health record design, emergency medicine training, air traffic control, workload, and human-automation teaming. She received her M.S. in Human Factors and Industrial/Organizational Psychology from Wright State University in Dayton, Ohio.

Katie Ernst is a human factors engineer with Applied Decision Science. Katie leverages her experience in military operations with methods from human factors and cognitive systems engineering to analyze and design for both military and healthcare applications. Prior to joining Applied Decision Science, she served 13 years in the U.S. Air Force in both active and reserve capacities. She received her M.S. in Industrial Engineering from Purdue University.

Laura G. Militello is co-founder and Chief Executive Officer at Applied Decision Science, LLC, a research and development company that studies decision making in complex environments. She is a recognized leader in the cognitive engineering and naturalistic decision-making community. Ms. Militello contributed to the development of early cognitive task analysis methods and coauthored a textbook on the topic (Hoffman & Militello 2009). She has applied cognitive task analysis methods to the design of technologies and training to support combat search and rescue piloting, air campaign planning, weapons directing, critical care nursing, and many other complex systems. She received her M.A. in Experimental Psychology and Human Factors from the University of Dayton.

Cover photo courtesy of the U.S. Army

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical information documents are available for download from the Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command