9-2020

Operationalized Intent for Improving Coordination in Human-Agent Teams

Michael F. Schneider

OPERATIONALIZED INTENT FOR IMPROVING COORDINATION IN HUMAN-AGENT TEAMS

DISSERTATION

Michael F. Schneider, USAF

AFIT-ENV-DS-20-S-074

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENV-DS-20-S-074

OPERATIONALIZED INTENT FOR IMPROVING COORDINATION IN HUMAN-AGENT TEAMS

DISSERTATION

Presented to the Faculty

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

Michael F. Schneider, BMEE, MS

Civilian, USAF

September 2020

AFIT-ENV-DS-20-S-074

OPERATIONALIZED INTENT FOR IMPROVING COORDINATION IN HUMAN-AGENT TEAMS

Michael F. Schneider, BMEE, MS

Civilian, USAF

Committee Membership:

Dr. Michael Miller
Chair

Dr. Gilbert Peterson
Member

Dr. Thomas Ford
Member

Dr. David Jacques
Member

ADEDEJI B. BADIRU, PhD
Dean, Graduate School of Engineering and Management

AFIT-ENV-DS-20-S-074

Abstract

With the increasing integration of artificial intelligent agents (AIAs) into human-agent teams (HATs), research into coordination mechanisms is needed to ensure members perform fluidly as a coordinated team. Research on coordination mechanisms in HATs is largely focused on AIAs providing information to humans to coordinate better (i.e. coordination from the AIA to the human). We focus on the complement where AIAs can understand the operator to better synchronize their actions with the operator (i.e. from the human to the AIA). We focus on AIA estimation of operator intent. We propose the Operationalized Intent framework which describes a portion of the operator's mental model in a manner relevant to "how" the team should perform. The core of Operationalized Intent is a quality goal hierarchy and an execution constraint list. Designing a quality goal hierarchy entails understanding the domain, the operators, and the AIAs. By extending established cognitive systems engineering analyses we developed a method to define the quality goals and capture the situations that influence their prioritization. Through a synthesis of mental model evaluation techniques, we defined and executed a process for designing human studies of intent. The resulting human-in-the-loop study demonstrates the feasibility of estimating Operationalized Intent. Using an existing label ranking algorithm, intent estimation is demonstrated with heuristically acceptable accuracy. This research establishes a path for designing intent aware multi-agent systems to enhance the implicit coordination of human-agent teams.

Acknowledgements

To my Leaders

Thank you for the opportunity to dig deep.

To my Guides

Dr. Mike Miller, my Committee Members, and Dr. John McGuirl

Thank you for helping me find the way.

To my Wingmen

Thank you for running with me to finish this race.

To my Support

My wife, and my five amazing children

Thank you for gifting me the space to accomplish this work.

From the days you pushed me into the office to work, to the nights you dragged me out to sleep.

To my Inspiration

My father and my mother

Thank you for the vision.

To my God

For the many miracles that made this possible.

Table of Contents

Page

List of Figures

List of Tables

Page

OPERATIONALIZED INTENT FOR IMPROVING COORDINATION IN HUMAN-AGENT TEAMS

I. Introduction

Overview

Teams are the mechanism by which complex work is performed. With the increasing integration of artificial intelligent agents (AIAs) into human-agent teams (HATs), research into coordination mechanisms is needed to ensure members perform fluidly as a coordinated team. These HATs are composed of multi-agent systems and multi-operator crews. Observation of current Air Force weapon systems indicates that HATs are already among us. An MQ-9 Reaper system, for example, includes autopilot, auto-router, sensor tracking, and health monitoring AIAs, to name only a few, managed by a pilot and sensor operator.

Coordination is a cyclical process critical to synchronized team action in human-human teams. It is no less true of coordination in human-agent teams. Coordination mechanisms in HATs are an area of active research largely focused on methods for AIAs to provide more information to allow humans to coordinate better with the AIAs (i.e. coordination from the AIA to the human). My research focuses on the complementary path where the AIAs can understand the operator and modify their own behavior to better synchronize with the operator (i.e. from the human to the AIA).

The challenges of coordination are multitudinous, so this research focuses specifically on AIA prediction of operator intent. Through rigorous analysis to characterize intent, we found that there has been much work focused on "what" the

operator intends to do, but remarkably little on "how" the operator intends the work to be done. While current AIAs are understood to be narrow in their scope, their breadth of control or trade space of action is growing. This expansion provides them more options in "how" they accomplish their task. The current research posits that an AIA which can understand "how" the operator intends to execute can better synchronize their actions with the operator. We established a theoretical framework named Operationalized Intent, which captures "how"-focused intent in a manner relevant to operators and AIAs. The core of Operationalized Intent is a *quality goal hierarchy* and an *execution constraint list*. To further explore the utility of this framework we focused on the design, study, and estimation of the quality goal hierarchy.

Designing an Operationalized Intent quality goal hierarchy entails understanding the domain, the operators, and the AIAs, which represent the three legs of the cognitive triad. By extending established cognitive systems engineering analyses we developed a method to define the quality goals and capture the situations that influence their prioritization. Through a synthesis of mental model evaluation techniques we defined and executed a process for designing human studies of intent.

This study produced a corpus of data which was used to investigate the ability of an AIA to estimate Operationalized Intent. Ultimately, we demonstrate that it is potentially feasible to design HATs in which the AIAs have a real-time model of the operator's intent to support team coordination.

Background

The current state of the art for human intent estimation are associate systems (Geddes & Buchler, 2012). Other efforts to estimate intent narrowly focus on specific tasks and do not holistically address operator intent. Born out of artificial intelligence advances in expert systems, Associate systems use Bayesian graph structures to reason about overall goals and methods of achieving goals. By observing the available system and user interaction data, associate systems develop an understanding of the team's situation and takes actions to further the goals it has identified. Associate systems have been demonstrated in research settings on aircraft (Banks & Lizza, 1991), helicopters (Miller & Hannen, 1999), command and control (Geddes & Buchler, 2012), and cyber defense systems (Huber & Marvel, 2016).

There are several limitations to associate systems. The system's model of intent is embedded in the graph structure which is opaque to the user and to other artificial agents (Geddes, 1989). Associate systems are purpose-built expert systems which entail a testing verification and sustainment burden which is sensitive to both overall system design changes and operational changes. There is an implicit design assumption that the operators wish to achieve their goals according to the methods used to create the graph structures. While associates can be networked with other artificial agents, they do not share their cognition, only the identified actions which should be taken. Associate systems serve as centralized generalist assistants for operators. However, the fact that their behavior is relatively opaque to the operator impedes them from truly coordinating with the operator. This is due to the fact it is, at best, difficult to direct and update the associates when their estimate of intent differs from the operator's true intent

Problem Statement

How can the effectivity improvements that human-human teams gain through implicit coordination be leveraged by a collage of AI agents in a human agent team?

Research Questions

The top level research question is: can a shared mental model be used to estimate the intent of an operator? This question is decomposed into the following research questions:

1. What is the framework for designing a shared mental model of operator intent which can be effectively implemented in human-agent teams? This framework should identify constraints and considerations from psychology, human factors, software, and systems design to provide a basis for designing an intent framework.
2. What is the process for applying Operationalized Intent to a domain? This includes the existing techniques that can be used to develop an intent model and identify relevant data elements needed to estimate intent along with the methods to study intent.
3. Do trial disturbances change the operator's intent in observable ways? The study trials are intended to shift the operator's elicited intent truth data over the course of the trial. If the disturbances do not shift intent, then intent cannot be estimated from the data.
4. Is elicited intent cohesive enough between operators, across situations, to be estimated in a generalizable manner? If the intent mental model is not a stable phenomenon with respect to the situation, it cannot be estimated situationally for a

generalized operator role. Instead it will be necessary to develop operator-specific intent estimation methods.

5. How accurately can an intent estimate be made from situational data as compared to the elicited intent model of an operator in a known and bounded context based on implicit communication and minimal explicit communication? Accuracy is judged by a modified Spearman Footrule metric (Diaconis & Graham, 1977). The situational data is gathered during a human-in-the-loop study.

Assumptions/Limitations

Operationalized Intent presupposes a human-agent team in which there are multiple agents with an ability to observe the environment commensurate with the operators. It is assumed that the operators will communicate explicitly with the agents regarding "what" is to be accomplished, but the agents will utilize their observations to estimate "how" they are to perform their activities. The operator is assumed to be trained, motivated, and earnest in the execution of the mission. Operationalized Intent seeks to represent the operator's priorities, whether they are appropriate or inappropriate for the situation. Thus, the operator is assumed to be properly in control of the overall system. An agent that critiques the operator based on the intent estimate is beyond the scope of Operationalized Intent. Operationalized Intent is expected to be convergent similar to the observed effect of other shared mental models (Fincannon, Keebler, Jentsch, Phillips, & Evans, 2011; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). Similarly, it is assumed that the operator's intent with respect to "how" tasks are to be performed is

stable in a stable task situation, i.e. in the absence of some disturbance to the situation. Thus as, discussed later, the intent estimate is valid for the duration of the situation.

Based on the literature, our research postulates that intent awareness via Operationalized Intent is highly likely to provide operational benefit to the end user. There is also likely utility in an intent estimate service for other artificial agents. The accuracy observed in this research is assessed heuristically, due to a lack integrated systems for effectivity evaluation. Operationalized Intent assumes that artificial agents have a decision making process which determines how they accomplish their task and that those paths produce meaningfully different results. It also assumes that making the decision regarding path selection requires a nuanced understanding of context. We assume that the intent estimate can be related to the situational data currently available in the system. The study assumes that the synthetic task environment is of sufficient fidelity that the results can be generalized to naturalistic environments. This research establishes if the concept of Operationalized Intent merits further investigation.

Research Method Summary

The overall research method sought to reach an initial validation point to demonstrate the potential of the underlying concept. As such we can divide the method into four phases.

Understanding: To address the research problem demands a thorough understanding of communication, coordination, and intent. We must comprehend the elements that underpin these crucial, yet frequently vaguely employed, terms and their effects on HAT performance. Through a synthesis of 50 years of research across robotics,

computer science, cybernetics, psychology, philosophy, neurology, human factors, and cognitive engineering, the basis for operationalizing intent is established.

Theorizing: From this rigorous grounding we develop the theoretical framework of Operationalized Intent and its application to HATs. This includes a method for development of intent models, and identification, capture, and representation of situational data via cognitive systems analysis. The intent model is composed of the *quality goal hierarchy* and the *execution constraint list*. With the end goal of estimating the intent, the intent models also must be characterized in a computationally estimable form.

Studying: Examining Operationalized Intent requires fastidiously designed trials. Combining multiple methods of mental model assessment with naturalistic decision making study techniques, we craft a method of situating operators to investigate their intent. We demonstrate this method through a human-in the-loop study with trained and experienced operators in a synthetic task environment (STE). The observed intent models are analyzed to demonstrate the effectiveness of Operationalized Intent in tactical scenarios.

Estimating: Leveraging the study data, an initial attempt to estimate intent and assess the accuracy of estimation is undertaken. Critical features are explored to assess their impacts and, ultimately, judge the estimation accuracy against the similarity of responses provided by study participants themselves.

It is noteworthy that these four phases do not align directly with the following four chapters. Aspects of the understanding phase are in Chapters II and III, the theorizing phase is presented in Chapters III, IV, and V while the studying phase is

presented in Chapters IV and V. The method for understanding Operationalized Intent, referred to as the Domain Application Process, is then summarized in Chapter VI. Due to the exploratory nature of this research, maximum use was made of available technology and experience. The Vigilant Spirit Control Station (VSCS) for Unmanned Aerial System (UAS) and supporting simulation software was readily available as an STE. Experienced, MQ-9 operators from the Michigan Air National Guard provided a baseline of expertise to study intent. Finally, the Label Ranking Random Forest (LRRF) algorithm was selected as a viable artificial intelligence technique to estimate intent as it is an open source estimator which had demonstrated superior performance on similar problems.

Dissertation Preview

This dissertation follows the academic format of compiling journal papers which encompass the research effort. As such there is overlap and review of concepts and conclusions in each of the following chapters along with some style and emphasis differences focused on the target publication audience. Chapter II reviews coordination and motivates the case for enabling implicit communication in HATs via intent. An analysis of intent and intent estimation in Chapter III provides a basis for the Operationalized Intent theory presented at the end of the paper. Chapter IV details the method for applying Operationalized Intent to develop a quality goal intent model, the human subjects study, and an analysis of the intent results. The method for capturing the situational data and the estimation of intent from that situational data is covered in Chapter V. While each of these chapters has their own conclusion section, Chapter VI concludes the dissertation with the contributions and future work.

II. Exploring the Impact of Coordination in Human-Agent Teams

Chapter Overview

This chapter lays out the synthesis of communication and coordination for the **Understanding** phase of the research method which demonstrate the utility of the research to improve human-agent team effectiveness.

Teaming is the means by which cognitively complex work is rapidly executed by multiple entities. As Artificial Intelligent Agents (AIAs) participate in increasingly complex cognitive work, they hold the promise of moving beyond being strictly a tool to becoming effective members of Human-Agent Teams. Coordination has been identified as the critical process that enables effective teams and is required to achieve the vision of a tight coupling between teams of humans and AIAs. This paper presents a characterization of coordination on the axes of content, types, and cost. This characterization is grounded in the human and AIA literature and is evaluated to extract design implications for Human-Agent Teams. This examination discusses the mechanisms, moderators, and models employed within human-agent teams, and identifies potential AIA design improvements to support coordination.

Introduction

Artificial intelligence systems, including those based on machine learning, have the potential to provide useful effort that can moderate human operator workload (Kaber & Endsley, 2007), enhance situation awareness (J. C. Gorman, Cooke, & Winner, 2006), and improve team performance outcomes (Kaber, Riley, Tan, & Endsley, 2001). However, the vision of systems in which human and machine intelligence are tightly

coupled to fully leverage their combined capabilities, as discussed by Licklider nearly 60 years ago, remains elusive (Licklider, 1960). In fact, a significant proportion of research in human-machine teaming continues to explore the interaction between a human and an artificial entity, rather than the integrated teams of humans and machines proposed in this landmark article. Recent performance increases and cost reductions in sensing, data transport, processing power, storage cost, and algorithm design (Brundage, 2016) are enabling Artificial Intelligent Agents (AIAs). The resulting AIAs are capable of sensing their environment, applying this information to support reasoning, and utilizing actuators to influence not only their environment, but other intelligent artificial agents, and human teammates (Weiss, 2013).

The cognitive functions that AIAs are addressing are increasingly abstract (Hare & Coghill, 2016) and dimensionally complex (Hubert et al., 2017). However, general artificial intelligence is not considered imminent (Goertzel, 2014). Current AIAs are assumed to be weak or narrow artificial intelligent agents (Kurzweil, 2005) that, while proficient in their specific task, lack generalized awareness and cognition regarding the world beyond their design. Due to the narrow focus of each AIA, systems are often composed of multiple AIAs to support higher-level goals (Weiss, 2013). The resulting multi-agent systems of AIAs have a large number of state and action spaces, allowing them to respond to the greater context, even if individual AIA capabilities remain narrow. Further, these systems are limited to tasks that are stable over time, have clear and measurable goals and can be characterized by a clear mapping of inputs to outputs (Brynjolfsson & Mitchell, 2017). Thus, successful deployment of AIAs in most complex, real-world environments will require these AIAs to be teamed with one or more humans

who are capable of innovating the application of these systems to solve more abstract challenges (Mercado et al., 2016; Rosenberg, 1982).

Following the design philosophy of adapting artificial systems to support naturalistic human interaction, designing AIAs to support teamwork requires that we understand the aspects of teamwork in human teams which are most important to enable in AIAs. A critical feature of human teams is how communication is employed to build the team cognition necessary to execute interdependent tasks (Salas, Cooke, & Rosen, 2008). We understand that sharing the cognitive load to enable coordination with team members has a cost (Woods & Hollnagel, 2006). Therefore, it is incumbent upon us, as system designers, to clearly understand and design AIAs to support appropriate cooperation within human-agent teams (Klein, Woods, Bradshaw, Hoffman, & Feltovich, 2004). In this research, we assume that the coordination observed in human-human teams is the greatest performance that is currently achievable by human operators in complex situations. Taking that as a design constraint for human-agent teams, we focus on how to develop AIAs designed to work as a member of these teams as opposed to designing systems that degrade human coordination to adapt to the limitations of AIAs. A necessary result of this perspective is the assumption that human agent coordination can eventually achieve the same performance as current human team members. Without redesigning human team member coordination, the AIAs must adapt to the available human coordination paradigm if we are to achieve high performing human-agent teams.

Thus, the goal of the current research is to review and clearly define team coordination as it relates to human-agent teams. This paper begins by unambiguously placing human-agent teams in context. We define and characterize coordination in

human-agent teams, including a) content, b) type, and c) cost. We then discuss the means for designers to improve the coordination of human-agent teams.

Clarifying the Design Goal of Human-Agent Teams

Teams are required to accomplish any task of sufficient complexity or criticality that require greater knowledge, skill, ability, work, or redundancy than a single operator can provide in the timeframe (Cooke, Gorman, Myers, & Duran, 2013; Salas, Dickinson, Converse, & Tannenbaum, 1992). In these environments, teams are formed with a sufficient number and diversity of teammates to provide adequate cognitive and physical capacity to overcome the complexity and criticality challenges. The team members not only bring diversified knowledge, skills, and abilities to the team, but train to work together interdependently (Delise, Gorman, Brooks, Rentsch, & Steele-Johnson, 2010). The result is a team comprised of individuals with varied perspectives on the task, which are derived from each operator's mental models of the task at hand, as well as their mental models of their teammate's ability to respond to task demands (van den Bossche, Gijselaers, Segers, Woltjer, & Kirschner, 2011).

High performing teams are thought to be comprised of diverse members who are committed to increasing their performance towards common outcomes. The team members collectively possess the skills necessary to address the task at hand, have the interpersonal skills (e.g., social sensitivity, emotional engagement, and communication patterns) necessary to perform as a team, and have the training to understand when, and willingness, to play specific roles within a team. At a minimum, these roles include creator, leader, and participant (Cheruvelil et al., 2014). As such, high performing teams

will select talented (Noe, Mcconnell Dachner, Saxton, & Keeton, 2011) and adaptable operators (A. Cox, 2017) from diverse disciplines (Kearney, Gebert, & Voelpel, 2009) and train them intensively (Delise et al., 2010) to focus on clear objectives (McComb, Green, & Compton, 1999). These team members take part in planning, execution and feedback processes, often referred to as transition, action, and interpersonal processes (Marks, Mathieu, & Zaccaro, 2001).

The focus of high performing teams often extends beyond the team members themselves to include examination and modification of policies and processes that impact their performance (Dickson, Singh, Cheung, Wyatt, & Nugent, 2009), creation of domain-specific vocabulary and gestures (Woods & Hollnagel, 2006), as well as, customization of supporting hardware and software systems (A. Cox & Szajnfarber, 2018). When employing supporting hardware and software, these teams often utilize operator creativity to take advantage of detailed control of the systems to extend the software and hardware functions beyond the designed system capacity with minimal design changes (Jacques & Strouble, 2010). As a result, they are able to adapt the system at a pace that outstrips the system development cycle (A. Cox & Szajnfarber, 2018).

Teams of AIAs certainly can possess knowledge and skills which extend beyond those of the human operator and thus may be desirable members of future high performing teams. However, the ability to coordinate individual adaptations towards a common goal, which is a capability that is often lacking in multi-agent systems, appears to be a key attribute for members of high performing teams. Thus, it is important that we decide whether to view multi-agent systems as adaptable hardware and software systems to be modified by the human team members or whether we seek to develop true human-

agent teams in which the AIAs are capable of coordinating their behavior with the team in response to the environment and the operators. The latter will require artificial agents that are capable of adapting to the needs of the human, as well as the environment, rapidly and reliably (Sycara & Lewis, 2004). While integrated user interfaces and agent transparency improve the operator's knowledge of the agents, permitting improved coordination (Stowers et al., 2016), the agents typically lack insight into the operator's understanding of the situation and task. Therefore, current artificial agents require explicit communication to coordinate (Klein et al., 2004) with their human teammates, typically during the mission, and thus are unable to leverage the less explicit forms of communication and coordination applied in high performing human-human teams(J. Y. C. Chen et al., 2018). As a result, there are frequent references within the human-agent teaming literature to implicit communication or coordination and intent inference (Espinosa, Lerch, & Kraut, 2004; Riley, 1989). However, these terms are often imprecisely defined and if we seek to design AIAs for well-coordinated Human-Agent Teams, we must understand coordination.

Content of Coordination

Salas and colleagues note that a critical feature of human teams is the communication employed to accomplish the team cognition necessary to execute interdependent tasks (Salas et al., 2008). We will thus refer to coordination as a cyclical communication process, verbal or nonverbal, which enables synchronized actions of teammates who are working on interdependent tasks. Based on this definition, while communication which supports coordination may occur within transition or interpersonal

processes, coordination occurs during action processes. We first attempt to characterize the content of coordination by reviewing the literature and proposing a general taxonomy of the content of coordination.

Several authors within the human factors, cognitive science, and artificial intelligence literature have addressed coordination. Klein and colleagues lay out the precursors to coordination and describe the mechanics of what they term the "Choreography of Joint Activity" in which parties signal changes in the phase of activity using coordination devices (Klein & Bradshaw, 2005). They examine the mechanics of how humans execute the coordination cycle, but there is little discussion of the content of coordination, beyond the need for planning the phases and the need to signal changes in phase. Malone and Crowston developed coordination theory with a focus on the types of dependencies (e.g. resources, task assignment, simultaneity, etc.) that the team is managing (Malone & Crowston, 1994). Their taxonomy addresses the foci of coordination and the extant processes particular to each of these foci. Within this body of work, Crowston develops a model of coordination activity using four language constructs: Information, Requests, Information Requests, and Proposed Actions (Crowston, 1991). Peterson and Bailey investigated air traffic controllers and developed a domain-specific taxonomy of topics as well as what they termed "communication formats." These communication formats include question, answer, statement, command, and command answer (Peterson & Bailey, 2001). Recent dynamic modeling research divided coordination into information, negotiation, and feedback which was applied to the specific human-human team paradigm being studied (Jamie C. Gorman, Amazeen, & Cooke, 2010). Table 1 summarizes this discussion of coordination categories.

Table 1. Coordination categories described in the literature.

Domain	Concepts	Reference
Human-Human Teaming (Medical)	Information Negotiation Feedback	(Jamie C. Gorman et al., 2010)
Management Science	Information (Declarative) Request (Imperative) Information Request (Interrogative) Proposed Action, Approved by Another (Performative)	(Crowston, 1991)
Human-Human Teaming (Air Traffic Control)	Question Answer Statement Command Command Answer	(Peterson & Bailey, 2001)
Artificial Agent Language (KQML)	Basic Query Multi-response Response Generic Informational Capability-definition Networking	(Vaniya, Lad, & Bhavsar, 2011)
Artificial Agent Language (FIPA-ACL)	Information Information Request Negotiation Performing Action Error Handling	(Huget, 2014)

Within the agent design literature, speech act theory has provided the foundation for AIA communication languages like Knowledge Query Manipulation Language (KQML) and the Foundation for Intelligent Physical Agent's Agent Communication Language (FIPA-ACL) (Vaniya et al., 2011). Vaniya and colleagues characterize KQML as having language constructs of multi-response, response, generic informational, capability definition, and networking (Vaniya et al., 2011). The performatives of FIPA-ACL are characterized by Huget as passing information, requesting information, negotiating, performing actions, and error handling (Huget, 2014). These techniques are

heavily tailored to their domains, which do not explicitly cover the human-agent team. However, these studies provide insight since they pertain to human-human teams working with machines or the communication between AIAs.

One can also discuss coordination in terms of speech act theory (Searle, 1969). However, the theory is designed to handle any kind of natural language communication. Coordination in a human-agent team may seldom be expressed in natural language due to the user preference for direct input human-machine interfaces (HMIs) over linguistic methods (A. L. Cox, Cairns, Walton, & Lee, 2008; Noyes & Starr, 2011). Therefore, it is necessary to characterize the content of coordination in a human-agent team.

Coordination is fundamentally a type of communication and so it can be expressed in the Shannon-Weaver model of communication (Shannon, 1948). Although it is recognized that communication is clearly cyclic between teammates, applying the Shannon-Weaver model allows us to examine the types or categories of information that may be passed from the teammate who is behaving as a sender at any moment in time. The temporal frame for a coordination event is the time required for the formulation of the communication in the sender, the transmission of the communication through a medium to the receiver, and the receiver's comprehension of the communication. Drawing on the above sources for inspiration, it is clear that some elements of information are intended to direct the receiving agent's behavior while other elements do not. Further, some elements of information are intended to be negotiable between the agents while others are not. Table 2 provides a proposed list of characters which differentiate coordination content. These characters are arranged by whether the content

should be negotiated by the team and whether executable directions are provided by the sender. Below we define each of the taxons of content within Table 2.

Table 2. Coordination content characters structured by the sender's expectations of the receiver.

		Directive	
		No	**Yes**
Negotiable	**No**	Information	Commands
	Yes	Plans Responsibilities Expectations	Requests

Information is situated data that enhances the awareness of the receiver, including directing the receiver's attention. In this regard, the sender should express this information in a manner that is concisely comprehensible to the receiver. Information is not negotiable because the sender believes it to be true, to some confidence level. If the receiver does not believe the information to be true, they can respond with corrected information based on their beliefs. The sender does not expect the receiver to perform a specific action based on the information. However, it is expected that the receiver will react appropriately given the assumedly agreed upon information (Jamie C. Gorman et al., 2010).

Commands are non-negotiable directives that the sender expects the receiver to carry out as soon as possible. The receiver may acknowledge the command by providing an information response indicating acceptance, execution, or completion. Commands cover a spectrum of content from fully specifying the exact activities for execution, to

stating a broad goal, allowing the receiver to formulate a plan and decide implementation details (Peterson & Bailey, 2001).

Requests are negotiable directives consisting of actions or tasks geared to achieve goals. The sender expects the receiver to carry out the request, but the receiver may reject or delay the implementation of the request. The receiver is expected to acknowledge the request and an information response will be provided, signaling the sender the receiver's decision to act, or not act, on the request (Klein & Bradshaw, 2005).

The remaining three taxons are all negotiable and are not directive. As such the information within these taxons requires a common understanding between the sender and the receivers within the team. The first of these are plans.

Plans are future-looking sequences of actions or tasks geared to achieve goals. A team may have a library of plans developed prior to execution, which has the potential to reduce cognitive and communication load during the activity. Alternately, these plans may be developed dynamically during execution. These plans answer the question of "what", "when", and occasionally "how" actions should be taken by a team member to achieve the agreed objective. A directive to enact a specific plan is command content, not plan content, which references a plan shared between the teammates (Huget, 2014).

Responsibilities are negotiable assignments of authority and answer the question of "who" is to conduct each activity. It is possible to have multiple operators and AIAs responsible for the same task or with the same authority, in which case they should be backing each other up (Cummings, 2014).

Expectations are the manner and procedures that the sender desires the receiver to use during execution. When commands or plans do not fully specify "how" to execute a

task, the receiver relies upon shared expectations to determine the timing and actions which will produce the desired execution.

Having established a taxonomy for the content of coordination, we can now examine different types of coordination and the impacts of coordination on communication within teams and the cognitive capacity required to facilitate coordination.

Types of Coordination Mechanisms

Coordination studies have noted and investigated the difference and development of human team coordination and found that coordination behaviors can be identified as explicit or implicit (Boos, Kolbe, Kappeler, & Ellwart, 2011; Entin & Serfaty, 1999). These types have been mapped orthogonally to other characterizations of coordination (Kolbe, Burtscher, & Manser, 2013), and for the purposes of this discussion, we consider the content of coordination as independent from the implicit or explicit type of the coordination exchange. Explicit coordination behaviors or mechanisms are identified by many different terms (Bolici, Howison, & Crowston, 2016; Butchibabu, Sparano-Huiban, Sonenberg, & Shah, 2016; Espinosa et al., 2004), but explicit coordination refers to communication directly or solely focused on managing dependencies and synchronizing actions. Implicit coordination is described as dependency management without dedicated or purposeful communication regarding synchronization (Butchibabu et al., 2016). Viewed in contrast, explicit coordination involves communication for the single purpose of coordinating activity. Implicit coordination involves communication that is multipurpose, providing context, which can be interpreted to imply activities necessary

for coordination. This does not imply that no communication is occurring for implicit coordination, rather, that team cognition is enabling existing communication and perception to be extended and used for coordination as well as its original purpose (Espinosa et al., 2004).

To clarify the distinctions, we draw between communication and coordination with respect to explicit and implicit types we propose the following definitions.

Explicit Communication: purposeful exchange of a discrete message (e.g. Unmanned Aerial Vehicle (UAV) pilot tells the sensor operator “we are nearing the area of operations”).

Explicit Coordination: explicit communication the primary purpose of which is synchronization of action in a team (e.g. UAV pilot tells the sensor operator “verify we are approaching the river shown on the map”).

Implicit Communication: observable behaviors, other than explicit communication, that convey additional information (e.g. UAV Pilot leans forward and switches from flying waypoints to manually flying the aircraft).

Implicit Coordination: Explicit or implicit communication the primary purpose of which is other than synchronization of action. Implicit coordination may involve implicit communication (e.g., leaning forward) or explicit communication (e.g. hearing pilot say, “we are nearing the area of operations”).

The proposed relationship between implicit and explicit forms of communication and coordination are depicted in Figure 1. Implicit coordination occurs in response to implicit, as well as explicit, communication. Explicit coordination is strictly the result of explicit communication.

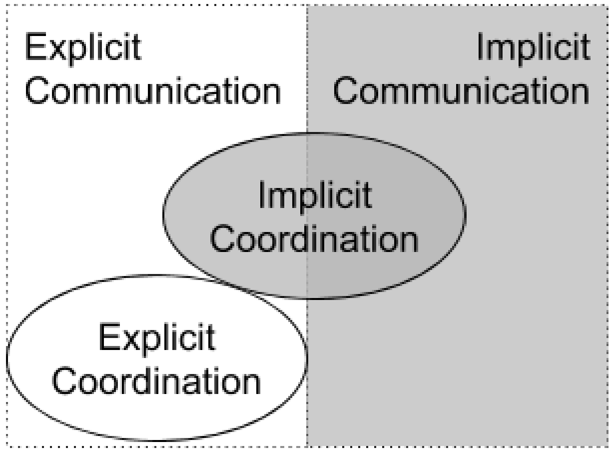


Figure 1. Explicit vs Implicit Communication and Coordination

Also important to this discussion is the fact that the communication can vary from abstract concepts to concrete data depending upon the desired level of control. This fact is recognized in system analysis models such as Rasmussen's ends-means hierarchies (Rasmussen, Pejtersen, & Goodstein, 1994); natural language processing models (Tomai & Forbus, 2009), and Geddes's operator intention model (Geddes, 1989). Therefore, while an explicit command pertaining to a high-level goal or abstraction layer provides explicit coordination; it also potentially implies numerous coordination activities.

Multiple studies have demonstrated that coordination shifts from highly explicit to increasing amounts of implicit coordination as teams become familiar with each other and the task (Butchibabu et al., 2016; Espinosa et al., 2004; Mathieu et al., 2000; Rico, Gibson, Sánchez-Manzanares, & Clark, 2019). Thus, the knowledge gained through experience with explicit coordination events enables implicit coordination by improving each team member's understanding of the team's response to a set of conditions. The literature indicates that humans adapt their coordination type depending on their perception of the team's cognition. If they observe that the team is managing

interdependencies well and the individual teammates perceive the team behaving consistent with their expectations, implicit mechanisms are more prevalent. Conversely, if the team is acting unsynchronized or an individual teammate does not perceive that the team behavior is matching their expectations, the teammate will likely fall back on explicit coordination mechanisms to improve interdependency management and refine team cognition. Importantly, this knowledge of coordinating activity is applied by each team member to select a set of coordinating actions. This is done without the sender issuing specific and highly detailed commands to each team member indicating the actions each team member should take to behave in a coordinated fashion. Coordinating activity is differentiated from general broadcast methods employed by many software agents that send their messages without regard to who the recipients are or what they need to synchronize their actions.

However, regardless of type or content, the organization, cognition, and action of synchronizing execution between team members requires cognitive resources. As system designers of Human-Agent Teams, it is imperative that we address the cost of coordination in the systems we design.

Cost of Coordination in Teamwork

Several authors have addressed the cost of coordination in teamwork. Klinger and Klein suggest that the cost of coordination increases with the addition of team members such that the marginal value of adding additional team members decreases with each increase in team size (Klinger & Klein, 1999). As this cost is readily apparent when viewing humans working in teams (Macmillan et al., 2004), it is important to understand

this cost and the effect that including additional AIAs within a team will have on coordination costs.

(Klein & Bradshaw, 2005; Schaeffer, 2009), classified these coordination costs into four categories: synchronization overhead, the time one entity spends waiting for another entity to complete a prerequisite task before beginning its task; communication overhead, the effort required to manage a handoff; redirection overhead, the time spent following an out of date plan after a new plan is signaled but before all entities understand the change; and diagnosis overhead, the additional burden spent diagnosing a problem that occurs as a result of interrelated activity. Klein and colleagues expand on these costs by listing the activities required to support coordination (Klein & Bradshaw, 2005). Among these activities are communication, monitoring, and feedback activities, some of which might be the responsibility of a frequent sender, such as a team leader, and some of which are clearly delegated to frequent receivers; i.e., the team members. Each of these works builds upon the work of Clark and Brenan who provide a list of costs for constructing common ground (Clark & Brenan, 1991). A review of this list clearly illustrates that some of these costs, such as formulation and production costs are born by the sender. Other costs, such as reception, understanding, and delay costs are primarily born by the receiver. Finally, many of the costs are shared between the entities; including start-up, directing attention, asynchrony (e.g., interruptions), change, display, fault, and repair costs.

Coordination can occur across the full spectrum from fully explicit (e.g., all coordination is explicit and messages must be complete) to fully implicit (e.g., a sender and receiver conduct activities without explicit communication). Further, the cost of any

given coordination exchange may vary significantly in magnitude for different entities within a team. As a result, we offer the construct in Table 3 to explore the cost of communication to the sender and receiver as it is required for coordination as a function of the interaction of team member role (i.e., sender or receiver) and communication type (i.e., implicit or explicit). As shown, each column represents a team member role and the two major rows indicate the extremes of explicit/implicit coordination type. Each of these two rows is further divided into “Activity Formulation”, “Communication Formulation/Understanding”, and “Communication Production/Reception. The “Activity Formulation” stage indicates which team member is responsible for interpreting the coordination content in the current context to create the sequence of activities the receiver is to conduct to maintain coordination. The following two rows represent the cognition necessary to support communication of these activities between the sender and receiver, as motivated by Clark and Brennen. The values within the cells of this matrix indicate an estimate of the relative cost born by either the sender or receiver. These relative costs, while likely continuous, are estimated as either full, partial or none within the cells of this table.

Table 3. Cost of communication required for coordination as a function of team member role and communication type.

		Sender Cost	Receiver Cost
Fully Explicit Coordination	**Activity Formulation**	Full	None
	Communication Formulation / Understanding	Full	Partial
	Communication Production / Reception	Full	Full
Fully Implicit Coordination	**Activity Formulation**	Partial	Partial
	Communication Formulation / Understanding	Partial (only critical)	Partial (only critical)
	Communication Production / Reception	Partial (only critical)	Partial (only critical)

In the case of fully explicit coordination, the sender, perhaps conceptualized as the team leader, must formulate all activities for each team member, formulate the communication of these activities, and produce this communication. The receivers must receive this communication, understand the activities to be performed and of course execute the activities. However, the receiver's cost of understanding the activities is likely reduced as the coordination must be fully specified. Importantly, at the extreme, fully explicit coordination requires the sender to formulate each team member's sequence of activities at the lowest possible, i.e. most detailed, level of control. As the coordination becomes more implicit the sender can begin to formulate and communicate at a more abstract level, for example by communicating the use of a plan. In this example, the receivers must now translate the plan based upon available coordinating information and

environmental context to formulate their own activities. Therefore, the sender incurs less of the cost of formulating these activities. At the extreme, the sender only formulates and communicates information critical to coordination, relying upon the receiver to leverage contextual information and non-explicit or non-coordinating communication to support activity generation and synchronizing.

Returning to the earlier discussion of the lists of the costs associated with coordination from the literature, each of these lists of costs focuses primarily upon the costs associated with communication or actions taken by other teammates. However, Clark and Brennan include the cost of directing attention, which is associated with the need for teammates to ensure that teammates are aware of environmental cues used to trigger changes in future activities (Clark & Brenan, 1991). It is rational to exclude the cost of perceiving environmental cues from the cost of coordination as this cost is also born by individuals interacting within an environment. However, plans or other coordinating information can include environmental events to trigger changes. Thus the activity required to observe these environmental events also plays a significant role in coordination and affects the coordination behavior of the team if not all team members are able to perceive the environmental cues. It is clear from this literature, however, that understanding of the team member's situation results in coordination costs.

Table 4 depicts the cost of coordination due to each team member's situation. Similar to Table 3, this table is divided into four quadrants with columns representing the costs incurred by the sender and receiver and rows representing fully explicit and fully implicit coordination.

Table 4. Cost of coordination due to understating team member's situation as a function of team member role and coordination type.

	Sender Cost			Receiver Cost		
Fully Explicit Coordination		Sender's Understanding	Receiver's Understanding		Sender's Understanding	Receiver's Understanding
	Sender's Situation	Full	None	Sender's Situation	None	None
	Receiver's Situation	Full	Full	Receiver's Situation	None	Full
Fully Implicit Coordination		Sender's Understanding	Receiver's Understanding		Sender's Understanding	Receiver's Understanding
	Sender's Situation	Full	Partial Cost	Sender's Situation	Partial Cost	Full
	Receiver's Situation	Full	Partial Cost	Receiver's Situation	Partial Cost	Full

Each quadrant is further divided into a two by two matrix. The columns of the matrix designate the costs associated with constructing their own understanding of the context or constructing their understanding of the other party's understanding of the situation. Further, the sender or receiver may need to understand their own situation as well as the receiver's situation. The fact that a team member must not only form their own understanding of their and their teammate's situation but project their team member's understanding of these situations may not be initially obvious. The need for the latter is discussed in an incident reviewed by Lee et al. in which a pilot, during a climb to a higher altitude, commanded the autopilot to descend to a lower altitude based on clearance from air traffic control (Lee, Hwang, & Leiden, 2015). The autopilot, instead of descending, entered a vertical speed hold mode and continued the climb as it was programmed to do in this specific situation. The pilot clearly understood his or her own situation and the receiver's situation. However, the pilot failed to track the receiver's

understanding of their own situation until the aircraft violated the air traffic control clearance. This resulted in a miscommunication that was fortunately resolved without incident. However, this example illustrates the need for the sender to understand their own and the receiver's situation, but also the value of the sender's projection of the receiver's understanding of their situation.

As shown in Table 4, when operating in a fully explicit mode, the sender is responsible for maintaining awareness of their own situation, the situation of their teammates who they are sending information to and their teammate's understanding of their own situation. When operating within an implicit mode, a portion of this responsibility shifts to the teammates. Although this shift reduces the cognitive load on the sender when they are coordinating with a single team member, this change in load increases when we consider the circumstance when the sender is the team leader and multiple receivers are included in the team.

To summarize the information in Table 3 and Table 4, explicit coordination, typically, consumes, predominantly, the cognitive resources of the sender who must formulate, as well as communicate an explicit command or request to each team member (Woods & Hollnagel, 2006). Implicit coordination distributes cognitive demand to the receivers. The receivers must then incur the obvious perceptual channel load, verbal, auditory, haptic, or visual, to gather information to serve as cues to the sender's situation. Further, they must execute cognitive processes to abstract the available information, understand the utility as it relates to the context, and formulate the appropriate coordinating activity (Endsley & Kiris, 1995). Implicit coordination reduces the real-time communication load on the team by multi-purposing a given communication event from a

sender to permit each teammate to select actions which will be coordinated with the actions of other teammates based on their understanding of each teammate's situation and their assessment of each team member's understanding of their current situation. It is important that as teams develop this knowledge, they often develop domain-specific terminology which encapsulates the description of system states with accepted plans (Woods & Hollnagel, 2006).

The utility of explicit communication to improve team effectiveness can be found throughout the teaming literature. For example, a meta-analysis of 150 team communication studies found that team communication positively influences performance (Lacerenza, Salas, Burke, Marlow, & Paoletti, 2017). Explicit communication, and perhaps, more importantly, communication to support explicit coordination, becomes an issue in time-pressured environments where the pace of execution and volume of information restricts the time available to explicitly communicate coordinating information (Entin & Serfaty, 1999). In such environments, the highest performing teams demonstrate well-coordinated behavior with limited or even no communication by governing their actions from team cognition that allow implicit coordination such that each individual can anticipate the actions of their teammates (Burke, Salas, Wilson-Donnelly, & Priest, 2004; Lacerenza et al., 2017; Stout, Cannon-Bowers, Salas, & Milanovich, 1999). For human-agent teams to improve their coordination during time-critical execution it is important to examine the mechanisms, moderators, and models for implicit as well as explicit coordination.

Designing Improved Coordination for Human-Agent Teams

Designers must understand their means to improve the process of coordination in human-agent teams. Since coordination is a process performed by members of a team, it can be influenced by system design. We propose that future AIA design must consider the three M's, as summarized in Table 5. These include coordination Mechanisms, Moderators of coordinating behavior, and Models used to coordinate.

The mechanisms of coordination are the methods used by the team to communicate (Okhuysen & Bechky, 2009) and have been discussed previously in the literature and earlier in this paper. As discussed, preplanning and debriefing tools (Stout et al., 1999) support transition and interpersonal processes which enable the construction of models to support coordination. Further, the literature discusses multiple mechanisms that permit the human to understand the information necessary to coordinate with an agent. These include common operating picture interfaces and shared information displays (Bolici et al., 2016), transparency focused interfaces (Mercado et al., 2016), and standardized callouts (Stanton et al., 2019). Although significant research has been conducted in building block technologies such as natural language understanding gesture recognition, and human state estimation, the literature appears to provide limited discussion of technologies which provide mechanisms which aid the AIAs in understanding the information necessary to coordinate with a human teammate.

In the context of human-agent teams care must be taken in designing the mechanisms to support the coordination of multiple narrow-focused AIAs. As discussed in the cost section, explicit communication between a sender and receivers can place a significant burden on the sender, particularly if the system is comprised of multiple

receivers. Assuming the human is performing as the sender and multiple AIAs serve as the receiver, this arrangement has the potential to place a significant coordination burden on the human. Further, the narrow-focused AIAs must be able to coordinate with each other and with the operators to avoid misunderstanding. The perspective of a narrow AIA may be much deeper in focus, but lack breadth and result in misinterpretation of a coordination mechanism with the operator. This observation leads one to consider the appropriate system architecture to provide the mechanism for coordination with humans within human-agent teams.

There are many factors that may moderate the effectivity of coordination behaviors and either enhance or degrade team performance. Examples include the ability and willingness of teammates to coordinate (Sukthankar, Shumaker, & Lewis, 2013), the flexibility of coordination mechanisms (Stachowski, Kaplan, & Waller, 2009), and the reliability or resilience of teammates in common and uncertain situations (Wohleber et al., 2016). While AIAs must be designed to be able to coordinate, the training program for operators must facilitate the human side of these moderators. Real-world operations frequently push the human-agent team beyond the designer's understanding which will necessitate adapting coordination to succeed, not destroying it with misunderstanding and unreliable behaviors during these unpredictable situations. The mechanisms must be designed to adapt to changes in these moderators.

Finally, there are the internal models which the individual team members employ to manage their coordination behavior. Examples include mental models (Gervits et al., 2020; Rico et al., 2019), transactive memory systems (Mesmer-Magnus, Niler, Plummer, Larson, & DeChurch, 2017), checklists, and scripts (Geddes, 1997). These models may

be of the goals, the work, the situation, or individual teammates, but they are critical to enabling implicit coordination. Artificial agents are known to predict the outcome of their own behavior. However, it is only by employing a model to estimate a future state in response to the interaction of their own and anticipated human activities, that AIAs can implicitly coordinate and reduce the cost of coordination on their human teammates. Specific sequences of actions may be highly procedural in nature for well-practiced task sequences, referred to as scripts by Geddes, or sequences of actions, which are adapted from previous experience (Geddes, 1989).

A specific subset of the teammate models include tracking the intent of an individual teammate to better predict their actions and needs (Ahmad et al., 2016; Y. N. Chen, Sun, Rudnicky, & Gershman, 2016; Holtzen, Zhao, Gao, Tenenbaum, & Zhu, 2016; Huber & Marvel, 2016; Kofler et al., 2015; McGhan, Nasir, & Atkins, 2015; Periverzov & Ilieş, 2015; Vered, Kaminka, & Biham, 2016). Based on the fact that the cost to the sender (e.g., human) is significantly higher when forced to employ explicit coordination with multiple receivers (e.g., AIAs within a multi-agent architecture), we propose that the exploration of explicit intent models for coordination, particularly in high performing teams, may be a fruitful area of future research.

Table 5. Mechanisms, Moderators, and Models to Improve Coordination in Human-Agent Teams

Concept	Examples	References
Mechanisms	Common operating picture, Transparency focused interfaces, Standardized callouts, Preplanning and debriefing tools	(Bolici et al., 2016; Mercado et al., 2016; Okhuysen & Bechky, 2009; Stanton et al., 2019; Stout et al., 1999)
Moderators	Ability and willingness to coordinate, Flexibility of coordination mechanisms, Reliability and resilience	(Stachowski et al., 2009; Sukthankar et al., 2013; Wohleber et al., 2016)
Models	Mental models, Transactive memory systems, Checklists or scripts,	(Geddes, 1997; Gervits et al., 2020; Mesmer-Magnus et al., 2017; Rico et al., 2019)

Conclusion

If we are to achieve the cyberneticist's vision of integrated human-agent teams (Licklider, 1960) it is important to design future multi-agent systems with AIAs which can coordinate effectively with human team members. Grounding our design in the coordination performance observed in human-human teams, this research focuses on design AIAs which utilize analogous coordination methods. We have provided a grounded framework for exploring and understanding coordination. By classifying the content and defining the types, we are able to provide further insight into the costs of coordination and to discuss the mechanisms, moderators, and models which are important for improving coordination in human-agent teams. As humans, we have been improving our coordination in teams for generations, with the advent of AIAs which can sense and react to the environment and its teammates, the time has come that our AIAs were designed to join the team.

III. Intent for Human-Agent Teams

Chapter Overview

This chapter details the **Understanding** phase related to intent and describes the **Theorizing** phase foundation for Operationalized Intent.

We seek to understand how intent might be integrated into future human-artificial intelligent agent (AIA) teams to improve coordination among team members. A brief review the use of intent for improving performance primarily within human-human teams is conducted to provide a better understanding of this term. This review clearly differentiates intent estimation from intent application, as well as the differentiation of "why", "what" and "how" based intent. A taxonomy of intent-based systems is then developed through a review of existing intent-based systems. Together these reviews show that intent has been modeled in a variety of ways without a cohesive understanding of intent and its different forms. Based upon these reviews and our understanding of multi-agent system architectures, we propose "Operationalized Intent" as a method of modeling intent regarding "how" the operators would like to execute the team's tasks. We propose that by embedding knowledge of how to execute within a multi-agent systems, the available AIAs within the system may perform their tasks in a manner that is more useful and synchronized with the operators and other AIAs within the system.

Introduction

Recent advances in deep learning, coupled with inexpensive sensors have reignited interest in artificial intelligence (Goodfellow, Bengio, & Courville, 2016). The resulting artificial intelligence systems when accompanied by appropriate sensors and

actuators will be referred to as Artificial Intelligent Agents (AIAs). These agents are able to sense their environment, apply previous knowledge to drive decision making, and take action in response to the decision (Weiss, 2013). While significant strides have been made within this field, the resulting systems remain narrow in focus and are generally unable to generalize their knowledge to adapt to uncommon or unforeseen circumstances (Brynjolfsson & Mitchell, 2017). To address the narrow focus of this automation, multi-agent systems are often explored where numerous AIAs are defined, each having a narrow focus, but functioning together to provide a broader effect (Franklin & Graesser, 1997; Weiss, 2013). Human-Agent Teaming, originally referred to as human-machine symbiosis (Licklider, 1960), is often discussed as a method for improving resilience and performance in unforeseen circumstances (M. Johnson, Vignati, & Duran, 2019). Teaming well trained, creative human operators with AIAs which are capable of performing specific tasks with beyond human performance can result in the high performing teams needed for complex environments (Driskell, Salas, & Driskell, 2018; Fiore & Wiltshire, 2016).

Coordinating communication between team members is crucial to team effectiveness in human-human teams (Eccles & Tenenbaum, 2007) and is a goal of design for human-agent teams (M. Johnson et al., 2019; Schneider, Miller, Jacques, Peterson, & Ford, n.d.). As is common in human-machine interaction, the exploration of human-agent teams leads to concerns regarding the efficacy of coordination between humans and artificial agents within these teams. This concern has led to an increased interest in communications among human team members (Marks et al., 2001) and the desire to understand whether aspects of this communication can be applied to improve

the efficacy of communication between humans and artificial agents within human-agent teams (J. Y. C. Chen et al., 2018; Lai, Chen, Zheng, & Khoo, 2020).

In the current research we examine leveraging intent for improving coordination in human-agent teams. The working definition of intent and intention is drawn from Bratman's work on the Belief, Desire, and Intention (BDI) model as: "relatively stable pro-attitudes that function as inputs to further practical reasoning in accordance with the two-level model of practical reasoning…" (Bratman, 1990). While understanding and communicating intent provides known benefits in human-human teams this topic appears to be gaining increased interest in human-agent teaming. Unfortunately, there appears to be little research or discussion within the human-agent teaming literature which attempts to provide an understanding of intent estimation or application. Therefore, the current research was undertaken to explore the components of intent, as discussed within the present literature, and its application within the human-agent teaming literature. The goal of this research is to determine whether a more nuanced view could be developed for the application of intent estimation systems. Based upon a discussion of this research and a literature review of systems employing intent, we then propose a taxonomy of intent estimation models. Finally, we propose a model for operationalizing intent research which we believe may be beneficial in near term human-agent teams, particularly multi-agent systems.

Intent to Improve Coordination

Among the aspects of coordination within human-human teaming is intent recognition and application. Intent recognition involves an individual's ability to

recognize another individual’s goals (Bonchek-Dokow & Kaminka, 2014). Recognizing this intent provides knowledge of likely future actions of team members. This projection can then be applied to predict a teammate’s actions and permitting an entity to plan its own activities without explicit communication. Intentions are future directed and serve three functions: they provide insight into the useful actions that may be taken, they guide the selection of actions, and they provide a hierarchical series of partial plans that guide deliberation (Bratman, 1990). Bratman observes that communicating intent is crucial to coordination between agents as it allows the sender and receiver a degree of mutual observability and predictability. Research has shown that human beings develop the ability to recognize intentions from observed actions between the ages of 9 and 15 months, even when the performer failed to achieve their goal (Meltzoff, 2005). Therefore, this ability appears to be developed early and serve as a building block of communication and coordination within human-human interaction. The ability to estimate intent is important in implicit coordination, which plays a significant role in the performance of expert teams (Eccles & Tenenbaum, 2004, 2007; Rico et al., 2019) .

Intent has been applied in an attempt to improve communication efficacy within the context of human-agent teaming at least since the 1980s (Geddes, 1989, 1997). This early research led to prototype systems which involved the development of an artificial associate to aid pilots or other operators (Banks & Lizza, 1991; Miller & Hannen, 1998). These early systems relied predominantly upon contextual information and operator inputs to the system to aid the recognition of human intent. This inference was then used to adjust the system’s response. Chris Miller has recently argued that the process of intent recognition and the automated application of implicit intent in human-AIA teams can be

fraught with error, leading to miscommunication between humans and their artificial agents (Miller, 2017). Therefore, he has argued for the use of high level, domain specific, language, i.e., plays, which aid the user in rapidly but explicitly conveying plans and intent to artificial agents within a system. Humans are known to develop domain specific languages to convey contextual information among human-human teammates in many complex environments (Woods & Hollnagel, 2006). Thus, the suggestion of the need for such domain specific languages to facilitate human-agent teams is not inconsistent with behaviors observed in human-human teams. However, as recognized by Geddes, commands or intent understood at one level of abstraction, implies behavior at lower levels of abstraction. Thus, it is reasonable to ask if specifying a play, particularly at a relatively high level of abstraction, may require knowledge of the situation and the user's intent to correctly execute the play.

AIAs must be a party to the shared situation context of the operators to estimate and apply the operator's intent. This context is comprised of the understanding of the environment, the individual operator's condition, and the operator's intentions. It is worth noting that these three elements are highly interdependent when applied to sense-making (Jeffery, Maes, & Bratton-Jeffery, 2005). The significance of an intention lacks specificity unless understood in the context of the immediate environment and the condition of the operator. In environments characterized by complex and networked information (e.g. command and control, unmanned systems, network defense, etc.), the AIAs are likely privy to at least as much environmental information as the operators, and potentially more than the operators maintain in working memory. The operator's condition, relative to workload, attention, alertness, etc., can be assessed through

behavioral and physiological measures (Albert & Tullis, 2013). Thus, the remaining piece of context the AIAs require is an understanding of intent, which must be relevant to their individual scope.

Understanding Intent

Child development research on intent recognition indicates that humans apply knowledge of human perception and motor function, observation of their teammate, contextual information and their knowledge of the context as input to the intent recognition process (Meltzoff, 2005). Much of this information likely comes from internal mental models, which are reconciled with observations of the teammate within the environment. Therefore, it is important to briefly review the structure and function of mental models as they relate to team communication and intent to gain an understanding of the types of intent and the communication of intent within teams.

Mental Models, Team Communication and Intent

We define a mental model as an internal representation of elements of an external reality (Johnson-Laird, 1980). However, elements of these models can be shared across team members (Mohammed, Ferzandi, & Hamilton, 2010). While these shared elements might be constructed by individuals observing a common environment, they are also informed by observing each other's behavior while operating within a common environment and through dialogue. The observations and communication can occur in training, prior to the operation, during the operation, or through after action operation analysis (Eccles & Tenenbaum, 2004; Entin & Serfaty, 1999; Jones, Ross, Lynam, Perez, & Leitch, 2011). As these mental models are applied to the use of man-made systems,

mental models permit us to understand the purpose, form, state, and function of a system, as well as predict the behavior of the system (Rouse, Cannon-Bowers, & Salas, 1992). The literature further differentiates shared mental models, i.e., models shared between dyads of individuals, and team mental models, i.e., models shared across an entire team (Lagan-Fox, Anglim, & Wilson, 2004) and shared mental models have been studied in depth in the context of teamwork (Langan-Fox, Code, & Langfield-Smith, 2000; Mathieu et al., 2000; Rouse et al., 1992). Team mental models also expand upon traditional mental models as they include information which enable team coordination rather than simply knowledge required to perform task (Eccles & Tenenbaum, 2004).

DeChurch and Mesmer-Magnus performed a meta-analysis of team cognition (DeChurch & Mesmer-Magnus, 2010). This analysis indicated that team behavior process and performance is improved when the teammates' mental models are similar in structure and content. In a recent update with 128 independent team cognition studies, these authors found that both compositional emergence (e.g. shared mental models) and team-centric content (e.g. teammate status/team process) had a significantly stronger correlation with team performance than compilational emergence (e.g. transactive memory systems) or task-centric content (e.g. work in progress) (Mesmer-Magnus et al., 2017). This indicates that shared mental models regarding team cognition improve team performance.

Importantly, common dialogue among team members or from team leadership often relies on shared mental model elements. For example, Entin and Serfaty found performance and teamwork scores improve when leaders are trained to periodically communicate key elements of their mental model to their team members (Entin &

Serfaty, 1999). Important to our present discussion, this research demonstrated an increase in implicit coordination mechanisms as the team members' mental models were synchronized with the mental model of the leader. This implies that once an operator's mental model is communicated explicitly to their team members, the team members apply this mental model representation to estimate the team leader's intent. However, humans do not require a single, monolithic, integrated mental model and will employ a variety of models during execution (Endsley, 1988). While efforts have been made to extend the concept of shared mental models to human-agent teams (Carpenter & Zachary, 2017; Howard & Cambria, 2013; Scheutz, DeLoach, & Adams, 2017; Yen, Fan, Sun, Hanratty, & Dumer, 2006), this effort has met with limited success. To provide further insight into mental models and intent, the types of information that are important within mental models, and thus useful in intent estimation, should be understood. Rouse proposed a taxonomy of training knowledge and skill requirements (Rouse, 1991) and extended it to include teaming knowledge and skills (Rouse et al., 1992). This taxonomy was proposed as a comprehensive basis to discuss types of information within mental models which are necessary to enable the team to execute successfully. Rouse's baseline training taxonomy is decomposed into three categories: System, Task, and Team. Each is further decomposed into three types of knowledge: "What", "How", and "Why." Each of these three types of knowledge has a continuum of levels ranging from detailed to abstract. Table 6 provides the overall structure of the taxonomy.

Table 6. Training Knowledge Taxonomy adapted from Rouse (Rouse et al., 1992)

Level	**"What"**	**"How"**	**"Why"**
System Knowledge			

Detailed	Characteristics of system elements	Functioning of system elements	Requirements fulfilled
Abstract	Temporal patterns of system response	Overall mechanism of system response	Physical principles/theories
Task Knowledge			
Detailed	Situations	Procedures	Operational basis
Abstract	Analogies	Methodologies	Mathematical principles/theories
Team Knowledge			
Detailed	Roles of team members	Functioning of team members	Requirements fulfilled
Abstract	Temporal patterns of team performance	Overall mechanism of team performance	Behavioral principles/theories

Overall, this taxonomy includes numerous elements which might be found within the current literature on team mental models. It is worth noting that the level of abstraction within this taxonomy is not clearly specified. Clarification of this level of abstraction and its relation to the types of knowledge deserves further discussion.

Use of "Why", "What" and "How" in Goal Hierarchies

Levels of abstraction are used in at least two distinctly different, but highly related, ways within the literature. A hierarchy within the cognitive systems literature is the Functional Abstraction Hierarchy (FAH), as presented by Rasmussen (Rasmussen et al., 1994). In this hierarchy, operator goals are decomposed into priorities, general functions, work processes, and finally material resources. Through the FAH, goals at one level are linked to the means (i.e., methods and apparatus) necessary to fulfill this goal at the lower levels. This tool is used to aid the understanding of the decision making within complex systems. Rasmussen proposes that "what" is present or to be accomplished at one level in this hierarchy can be observed by selection of a level. The layer above defines "why" one would accomplish the "what" and one layer below defines "how" the

item may be accomplished (Hollnagel & Woods, 2005). The FAH is not used to truly decompose system goals and functions but to aid the understanding between physical elements of the system and higher-level mental constructs of the system. Although it is not clear that all of the abstract components in Rouse's taxonomy are consistent with abstraction as applied in FAH, it is clear that at least some of the abstract concepts are common between both. For example, analogies are higher level mental constructs formed by the user rather than system functions. As Lind notes, FAH struggles to "differentiate teleological and causal reasoning representations" which means that the hierarchy of "why"-"what"-how struggles to be fully generalizable (Lind, 2003).

In an alternative approach, a goal hierarchy is formed in which achievement of the lower level goals contributes to the higher level goals (Endsley & Jones, 2012; Geddes, 1989; Humphrey & Adams, 2011; C. Johnson, Miller, Rusnock, & Jacques, 2020; M. Johnson et al., 2012). This approach somewhat mirrors that of a common robotics architecture, termed the subsumption architecture (Brooks, 1986). Originally developed by Brooks, the subsumption architecture requires that the robotic component be designed to achieve a lower level goal through completion of the full perceptual cycle and that this goal is embedded in a higher level component which embeds or subsumes this behavior with other components to form a higher level behavior necessary to achieve a higher level goal (Brooks, 1986). An interesting aspect of these goal hierarchies is they often enable multiple means to accomplish a goal at a given level. For example, Geddes discusses the fact that humans can achieve a goal utilizing one or more "plans", i.e., loosely associated steps which might be taken in some order to accomplish a goal, or "scripts", i.e., proscribed procedures which are followed in a specific sequence, to achieve a goal

(Geddes, 1989). This is similar to the Goals, Operators, Methods, and Selection Rules as proposed by Card, Moran and Newell (CMN-GOMS) model (Card, Moran, & Newell, 1983). CMN-GOMS also acknowledges that human operators commonly apply selection rules to select among multiple methods to achieve a goal, depending upon the situation in which they are operating. Therefore, similarly to the FAH, analyzing the goal structure at a select level, the goal at any level of analysis within the goal hierarchy specifies "what" is to be accomplished. An associated goal at the next higher level within the goal hierarchy specifies "why" the goal is to be accomplished. Selecting a means or the method to accomplish the goal, specifies "how" the goal is to be accomplished and this selection is dependent upon the present situation. Notice that in this goal structure, the underlying goals do not necessarily specify "how", but describe subservient goals which might be achieved to achieve the higher-level goal.

In discussing modeling of goals for human-agent teams, Sterling and Taveteer discuss functional goals, as well as quality goals (Sterling & Taveter, 2009) . While the functional goals correspond to "what" is to be achieved, quality goals pertain to the priority of various criteria and constraints necessary to understand "how" the functional goals are to be attained. For example, this text discusses the quality goals of performance, safety, and security, among other socially driven goals, for a palletizing robot. The perceived utility of the robot not only pertains to whether the robot completes its work, but also the way that it balances performance (e.g., pallets completed per unit time) with these other considerations given constraints in the environment.

To illustrate this concept through an example, we may have a mission to surveil a target using an aircraft, specifying "why" the human-agent team is engaged in operations.

An operator can lay out the "what" sub goals: find the target, establish positive identification, set up surveillance pattern, maintain custody of the target, and provide reports on target observations. However, when asked "how" those goals are accomplished, the answer is invariably: it depends. The situational considerations, including sensor selection, standoff distance, altitude, and airspeed, drive the selection rules for the individual tasks that need to be executed. For instance, if conducting surveillance of a capable adversary having the ability to detect the aircraft, one might select a high airspeed to maximize maneuverability and aid survivability together with altitudes and standoff distances which permit one to ensure sufficient detail in the target can be observed in a single pass. However, when observing a less capable adversary one might select lower airspeeds to minimize fuel burn and maximize time observing the target, together with altitudes and standoff distances just great enough to reduce the likelihood of visual or auditory detection.

From this review we propose a relationship between Rouse's types of knowledge and levels of abstraction as they relate to intent in human-agent teams as depicted in Figure 2. In this figure, the goal hierarchies to support human-AIA team design should consider "why" knowledge as specified by the top level functional goals, e.g. the mission. This goal is relatively stable, although high level mission goals can change. However, these changes are typically driven by external entities, such as a commander in a military context. "What" knowledge is defined by a plan or sequence of the functional sub goals that compose achievement of the "why" goal. As this knowledge specifies each task to be performed by each entity, it changes relatively rapidly. The "how" knowledge is driven by situational considerations and change based upon changes in the environment. In the

surveillance example, the capability of the adversary drives how the tasks are executed. The “how” knowledge drives changes in the method that is applied to accomplish the functional goals and therefore guides and constrain the functional goals. We propose that in a human-AIA team, this knowledge can be specified through the use of quality goals which drive the selection rules and thus define “how” the individual methods are executed to accomplish the functional goals. For the present discussion, it is important that “what” must be designed at a level in the hierarchy where the human is actively teaming with an AIA to accomplish shared goals to understand intent in human-AIA interaction.

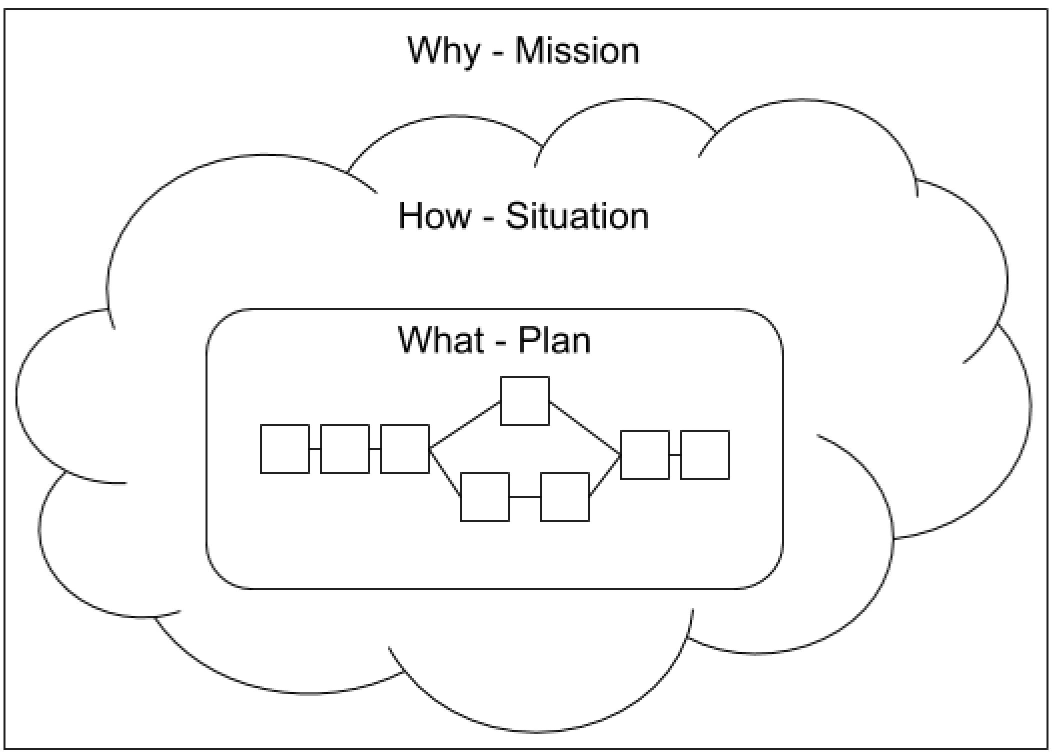


Figure 2. Relationship between “Why”-”What”-How based on Goal Decomposition

Intent Implications for Human Agent Teaming

Overall, the literature indicates that humans perform at least three stages of intent processing. The first of these stages involves the identification of whether an action performed by another individual has purpose or is useful in intent estimation. Studies in child development have indicated that infants will follow head movements of individuals without a blindfold, recognizing that this head movement provides information about the object another human perceives. However, if these infants understand the function of a blindfold, they are less likely to exhibit this same behavior when the other individual is blindfolded (Meltzoff, 2005). Apparently, the infants recognize that knowledge of the other individual's perception, as indicated by their head movement, provides goal driven information useful in the intent estimation. The infants recognize that with the blindfold the other individual is not engaged in the goal driven behavior necessary for intent estimation. The second stage involves the estimation of intent, wherein the individual applies information to estimate the goals or intent of another individual (Bonchek-Dokow & Kaminka, 2014). The third stage is the application of the intent estimate to aid decision making and drive the individual's action in response to the intent estimate.

Although this developmental research explored implicit communication of intent in which an individual observes the behavior of another to estimate intent to guide future action, as Chris Miller indicates, intent can be communicated implicitly or explicitly (Miller, 2017). In explicit communication of intent, the individual can communicate their intent, making the understanding of goal driven behavior and intent estimation unnecessary. In either of these situations, this intent can then be applied to drive decision making and future behavior of the observer.

If we are to envision the design of multi-agent systems where the AIAs leverage intent understanding to improve performance in human-agent teams one might expect that functional agents could apply estimates of intent to aid decision making and action selection. However, for the behavior of each of the functional agents to appear consistent and predictable to the human operator, each of the agents will likely need to leverage a similar intent estimate to drive decision making and action selection. Therefore, it is reasonable to employ an individual agent to determine operator intent. This agent can either interact with the human to obtain an explicit specification of intent, or identify purposeful human actions and implicitly estimate elements of operator intent.

As noted earlier, it may be desirable that certain elements of intent be determined through explicit communication while other elements are determined without explicit communication. Play calls, discussed by Chris Miller, provide a short-hand to explicitly specify a plan, which specifies the "what" information within the Rouse et al.'s knowledge taxonomy. The "why" information can also be specified explicitly if needed by the AIAs. However, as the "how" information is based on the situation, it may be reasonable to derive this information from knowledge of the environment, system, and observations of the human's actions.

Characterizing Intent

To gain further insight into implicit intent estimation, a literature review was conducted to understand how recently described AI-based systems attempt to utilize implicit intent. Recent studies of intent driven AIAs have focused on seven domains, including aviation, driving, human-robot interaction (HRI), surveillance, cybersecurity, and general human-machine interface (HMI). This literature includes a diverse array of

models of intent. Many focus on physical trajectory (e.g. lane change, meeting someone, activating a control, etc.) while others compare observations to an a priori knowledge base to identify intent (e.g. goal list, simulation model, plan goal graph). The models of intent developed are largely defined by the specific problems the research addresses. To discuss these intent estimation efforts further we propose characterizing them in a taxonomy, depicted in Figure 3. The taxonomy includes eight classification categories which are defined in the following paragraphs.

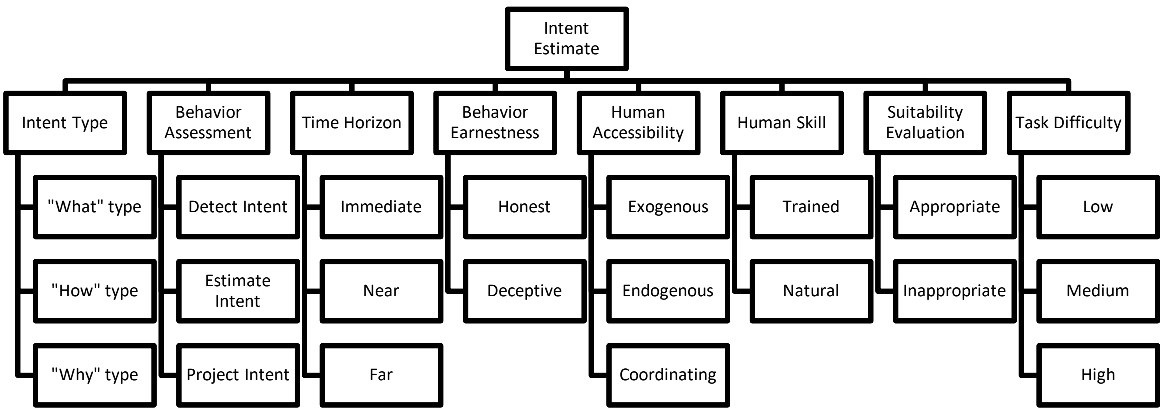


Figure 3. Intent Estimation Taxonomy

Intent Type defines the type of information which is estimated by the system. Three categories of intent type can be observed, including: 1) “What” estimates the end goal or next action to be pursued by the human; 2) “How” estimates the methods, constraints, and qualitative desires that constrain the execution of the human or the team; 3) “Why” estimates the driving purpose for taking an action or seeking a goal.

Behavior Assessment defines the explicit actions taken within the intent estimation system. Three categories of behavior assessment are present, including detection, estimation, and projection. Systems performing detection determine if the human is behaving in an intentional manner. This assessment can be conducted to

determine if a behavior should be included in further estimation efforts. Detection is applied by Bonchek-Dokow to evaluate intentionality of humans in an observed scene (Bonchek-Dokow & Kaminka, 2014). Systems performing estimation attempt to determine intent based on observations of human behavior. As an example Lee estimates pilot intent as behavior consistent with their control input (Lee et al., 2015). Systems performing projection estimate intent beyond the immediate intent. For example, Krozel estimates the flight path of an aircraft beyond the current maneuver (Krozel & Andrisani, 2008), thus projecting not only immediate intent but future intent.

Time Horizon defines the time scale of the intent estimate. Three categories were observed in literature review. Systems with an immediate time horizon attempts to estimate intent of the current action. For example, Vered's goal mirroring algorithm continuously revises the intention estimate (Vered et al., 2016) to understand the following step to be performed by the individual. Systems with a near time horizon estimates intent for some small sequence of actions. For example, McGahn's Markov Decision Process model provides estimates of multi-action tasks (McGhan et al., 2015). Systems employing far time horizons estimate a large sequence of actions. For example, the pilot's associate, as described by Banks and Lizza, re-plans entire missions based on changes in intent (Banks & Lizza, 1991).

Behavior Earnestness indicates the level of cooperation that the observed individual is assumed to exhibit during intent estimates. This taxon includes two categories, honest and deceptive. Systems which assume the observed individual is exhibiting honest behaviors, expect that every action the human makes is performed to accomplish a goal. For example, Kelley uses hidden Markov models which assume that

the observed humans are seeking goals (Kelley et al., 2008). Alternately, systems were discussed which assume that the observed individual is performing actions to maliciously manipulate the system to perform poorly or undesirably. For example, Cuppens uses intrusion scenarios to identify malicious cyber-attacks (Cuppens, Autrel, Miege, & Benferhat, 2002) wherein actions are performed to camouflage the true goal of the cyber-attack.

Human Accessibility indicates the inclusion of the human being observed. This taxon includes three categories: exogenous, endogenous, and coordinating. In exogenous systems, the human being observed is external to the human-agent team, permitting only immediate situation and behavior to be observed. For example, Huang infers the intent of other drivers on the road informed by road curvature (Huang, Liang, Zhao, Yu, & Geng, 2017). In Endogenous systems the human is internal to the human-agent team and system. Within these systems, team, historical or coordinative information is available to the agent in addition to situational and behavioral information. For example, the studies reviewed by Xing infer the intent of the driver of the vehicle which is teamed with the agents (Xing et al., 2019) which are seeking common goals. In coordinating systems the agent can explicitly coordinate with the human being observed to acquire information to assist in intent estimation. For example, in the rotorcraft associate systems, the agent can communicate directly with the pilot to obtain feedback or explicit intent information (Andes, 1997; Miller, 2017)

Human Skill refers to the assumed knowledge of the human being observed. This taxon includes two categories, including trained and natural. Systems assuming trained operators make the assumption that the humans have some experience, qualification, or

instruction specific to the system in which the human-agent team is embedded. The agents can confidently make assumptions about the mental models the human is employing and the procedures they are applying. For example, Banks and Lizza leverage the training of pilots to infer goals from behavior (Banks & Lizza, 1991). Systems assuming natural operators assume the human may have no training or experience with this specific system. Instead they only have domain knowledge relevant to the task. For example, Holtzen observes humans in a natural office setting accomplishing mundane tasks (Holtzen et al., 2016).

Suitability Evaluation refers to whether the system assumes the human is capable of accurately forming intent or if the system assumes the human may make errors, requiring that their intent be corrected based upon a situated standard of intent. For systems assuming appropriate intent the estimate of the human's intent assumes their intent is valid or correct for the situation and the system should apply this intent to adapt its behavior. For systems assuming inappropriate intent the system must determine whether the estimate of the human's intent is valid or invalid for the situation before selecting an action.

Task Difficulty refers to the amount of noise or dimensionality of the task the human agent team is accomplishing. Difficulty can be expressed as shown in Table 7.

Table 7. Dimensions of Task Difficulty

		Environment	
		Abstracted	**Complicated**
System	**Simple**	Low Task Difficulty	Medium Task Difficulty
	Complex	Medium Task Difficulty	High Task Difficulty

The literature and a few important characteristics of the systems discussed in this literature are summarized in Table 8. As shown in this table, the majority of these systems seek to estimate the "what" element of intent. Only two of these systems estimate intent to guide "how" they respond and only one focuses on inferring the "why" of intent. In addition to the domain and the type of intent, Table 8 briefly indicates the task performed by the system and whether the system is focused on determining the intent of a member of the team (e.g., is focused on endogenous intent) or the intent of an external human (e.g., is focused on exogenous intent). Additionally, an attempt is made to order these systems in terms of increasing difficulty of intent estimation.

Table 8. Overview of Intent Estimation Literature

Reference	Domain	Focus	Task	Intent Type
(Huber & Marvel, 2016)	HMI	Endogenous	Decision Support	What/How
(McGhan et al., 2015)	HRI	Exogenous	Trajectory prediction	What
(Ahmad et al., 2016)	HMI	Endogenous	Search engine selection	How
(Kelley et al., 2008)	Surveillance	Exogenous	Goal estimation	What
(Kofler et al., 2015)	HMI	Exogenous	Video classification	Why
(Lee et al., 2015)	Aviation	Endogenous	Anomaly detection	What
(Periverzov & Ilieş, 2015)	HMI	Endogenous	Clarify ambiguous inputs	What
(Vered et al., 2016)	HRI	Exogenous	Trajectory prediction	What
(Ahmad et al., 2016)	HMI	Endogenous	Clarify ambiguous inputs	What
(Bonchek-Dokow & Kaminka, 2014)	Surveillance	Exogenous	Goal estimation	What
(Y. N. Chen et al., 2016)	HMI	Endogenous	Clarify ambiguous inputs	What
(Cuppens et al., 2002)	Cyber	Exogenous	Malicious attack detection	What
(Holtzen et al., 2016)	HRI	Exogenous	Trajectory prediction	What
(Krozel & Andrisani, 2008)	Aviation	Exogenous	Trajectory prediction	What/How
(Huntsberger, 2011)	HRI	Endogenous	Collaboration	What/How
(Xing et al., 2019)	Driving	Endogenous	Lane Change	What
(Andes, 1997)	Aviation	Endogenous	Collaboration	What/How
(Banks & Lizza, 1991)	Aviation	Endogenous	Collaboration	What/How
(Ferguson & Allen, 2007)	HMI	Endogenous	Natural language interface	What
(Geddes, 1997)	Aviation	Endogenous	Collaboration	What/How
(Huang et al., 2017)	Driving	Exogenous	Trajectory prediction	What

This literature review reveals a few interesting aspects of existing intent estimation systems. First, it is obvious that some applications are attempting to determine the intent of individuals who are external to the system. Further, some systems are

attempting to understand the intent of individuals who are actively attempting to conceal their intent. In each of these systems, it is not reasonable to expect these individuals to engage in communication to explicitly convey their intent and therefore implicit intent estimation is necessary. Although the primary interest of the current research was to reduce the effort required to engage in fully explicit communication in human-agent teams, the fact that individuals may not engage in cooperative behavior to enable estimation of implicit coordination has the potential to increase the difficulty of forming robust estimates. Considerations such as the time horizon, variations in human skill, and the potential to form inappropriate intent all provide potential further complications to the estimation of implicit intent and its application.

The literature review further indicates that intent estimation has been investigated in many contexts of varying complexity, informing the desirable execution of an AIA. However, each of the extant systems attempt to estimate "what" type knowledge and only attempt to model "how" type knowledge through the actions the AIA decides to take. In the Jet Propulsion Laboratory study of a robot inferring astronaut intent, the "how" knowledge and estimation are embedded in the ever branching model of "what" they are doing and will likely do next (Huntsberger, 2011). Given our earlier discussion it is possible that the "how" type knowledge may be more stable and change in response to environmental and system variables in addition to human attitude. As these variables are likely more accessible to the AIA, these changes may improve the AIA's ability to characterize and respond to these changes. For example, an aircraft autopilot which understands how maneuvering impacts fuel efficiency must comprehend the importance of fuel efficiency versus deviation from the flight plan or other considerations. The

"what" of flying the aircraft is unchanged, but "how" it should be flown can have significant safety and mission impacts. These "how" intent considerations must be synchronized between AIAs and coordinated with the operators. Therefore, developing methods to address coordination mechanisms for "how" type intent may prove useful in developing complex multi-agent systems for human agent teams.

Operationalizing Intent

Given this background, we propose a new human-AIA coordination mechanism named Operationalized Intent. To apply this mechanism, we propose that if one can understand the priority an operator would apply to quality goals within a specific situational context, these goals and their priorities can be used to guide the behavior of a number of agents within a multi-agent system while reducing the need for coordination regarding "how" type intent. While the earlier surveillance example is an extreme, it is possible to see how understanding the tradeoffs between the operator's quality goals with respect to safety and data quality as well as the context of these two separate missions might have helped the system estimate how the mission should be conducted.

Modeling intent in a manner specific to an AIA is appropriate when addressing how that AIA understands intent for its tasks. However, we propose that when integrating a human-agent team with multiple AIAs and operators, intent must be coordinated between teammates and must therefore be explicitly modeled.

Model

The implications of this discussion to intent are that communicating situational context, i.e. "how" intent, reduces the appropriate task execution methods. Given the

limited capability of current AIAs to construct robust shared and team mental models, in systems where a human operator is teamed with one or more AIAs, it is likely that the "what" is to be accomplished should be explicitly communicated. However, we propose that explicit communication regarding "how" to perform the task could be reduced, reducing the burden on the operator, by modeling operator intent and engaging in implicit coordination of "how" to execute the appropriate tasks. It is important to create this model such that it can be shared with the operator and the agents, giving the operator a mechanism to communicate explicitly if required.

The core of Operationalized Intent is the explicit intent model. The intent model is a shared structure composed of two elements, an ordinal prioritized list of quality goals, and a list of execution constraint statuses. The quality goals and execution constraints are established during a design process, in coordination with operators and AIA designers, and do not change during execution. Quality goals are ranked hierarchically according to their relative importance to the operator and change throughout the course mission of execution. Execution constraints are assigned a status, normal, enforced, or overridden, which describes how each constraint is influencing AIA behavior during execution. The prioritized goal hierarchy and constraint status list form the intent estimate which is an immediate representation of the operator's internal state during execution. As Krozel (Krozel & Andrisani, 2008) demonstrated, intent estimates can be composited to extend the time horizon. A set of temporally sequential intent estimates define a published intent for an operator. Published intent is specific to a given operator in multi-operator human-agent teams.

Quality goals describe task execution relevant guidance. A useful quality goal complies with the following heuristics. First, it is relevant to the span of control of at least one task of one AIA in the system, such that the AIA is capable of changing its method due to changes in goal priority. Second, quality goals must be conceptually relevant to at least one operator in the system such that he or she can understand and communicate the priority of this goal. Third the operator must be able to differentiate each goal from other goals. Fourth, the AIA designers must be able to render the goal into a computational form such that the system can adapt to its priority in a continuous manner. Finally, the AIAs must be able to violate the guidance imposed by a quality goal given other considerations, enabling tradeoffs during system execution.

Execution constraints define some limit to the trade space of behaviors available to the AIAs. These execution constraints include both soft constraints, which can be violated in special circumstances and hard constraints which cannot be violated at any time. Describing the impact of execution constraints on AIA behavior drives the discrete statuses which can be communicated with the human operator. These include normal, enforced, and overridden. Normal status is the condition in which the constraint is not immediately limiting the behavior of the AIAs. Enforced status describes the condition in which the constraint is immediately limiting the behavior of the AIAs. A designated responsible AIA determines the normal or enforced status. Overridden is a contingency status which allows the operator to instruct all AIAs to ignore a soft constraint. For example, the aviation constraint “Do not descend below the minimum safe altitude”, must be overridden to land the aircraft.

The substantive difference between quality goals and execution constraints entails that a single statement cannot be both a goal and a constraint simultaneously. It can be difficult to determine where an execution expectation should fall, particularly when considering pervasive ones like safety. Consider the directive to “Operate Safely.” While this directive is often viewed as the top priority in safety critical systems, as a quality goal it struggles to be interpretable, differentiable, and it wouldn’t be expected to ever be violable. However, an execution constraint that stipulates “Do not operate above critical thresholds” provides actionable limits on the AIA’s behavior. This constraint does not provide any guidance on operating close to critical thresholds, only that they should not be breached. This is true even though operating near such a threshold increases the likelihood of failure, reducing the margin for error. When such a constraint is paired with a quality goal to “Avoid operation outside normal range”, the expectation about safe operation is usefully decomposed into a hard constraint that prevents unsafe actions by the AIAs, and a quality goal that promotes safe operation by providing guidance to generally avoid operating near constraint bounds without strictly limiting behavior. It is still possible to operate near critical thresholds given circumstances where other goals are of greater importance, but there is an expectation that the system will operate away from the edges of the envelope under conditions where the quality goal “Avoid operation outside normal range” is more important than other quality goals.

Characterization

To clarify the purpose and utility of Operationalized Intent we evaluate it according to the intent estimation taxonomy provided in Figure 3. With regard to the taxon “intent type”, Operationalized Intent addresses the “how” type of intent. AIAs are

focused on their tasks and operators should clearly, unambiguously, and directly (Shneiderman, Plaisant, Cohen, & Jacobs, 2010) control or command the team. The intent estimate provides AIAs with the added context, improving their ability to select appropriately among different options for accomplishing a designated task. With regard to the taxon “behavior assessment”, the operator is always assumed to be behaving intentionally. Operationalized Intent seeks to develop intent estimates in real time. Integrating Operationalized Intent into a planning process provides a projection of intent estimates into the future. The intent estimate is assumed to be valid for an immediate or short “time horizons.” While published intent might provide intent estimates at further time horizons, it is expected that the AIAs will not rely upon prognostications that extend significantly into the future. There is an implicit assumption in human-agent teaming that the operators are working with the AIAs to achieve the goals, by extension Operationalized Intent assumes operator honesty with respect to “behavior earnestness.” With regard to “human accessibility”, the implementation of Operationalized Intent gains efficiency and power if implemented in a coordinating manner. Operators should have visibility into the agent’s intent estimate and a method to resolve conflicts between the estimate and the operator’s true state should be provided. For the “human skill” taxon, the design of the intent model necessitates that the operators be trained. Operationalized Intent is focused on high performing teams which, by design, train together to improve performance during execution. With regard to “suitability evaluation”, Operationalized Intent is designed for application in natural environments which are notoriously noisy. Therefore, intent estimates focus on representing the operator’s state without judgment regarding the suitability of that intent to the situation. Finally, on “task difficulty”,

Operationalized Intent is targeted at complex systems operating in complicated environments. The utility of explicitly modeling “how” type intent arises with increasingly complicated environments which pose multiple dilemmas to the team’s joint cognitive abilities and demand an increased variety of action (Ashby, 1956) to address the multidimensional nature of the environment.

Operationalized Intent is most similar to the associate systems reviewed earlier (Andes, 1997; Banks & Lizza, 1991; Geddes, 1989). Critical differences include the fact that Operationalized Intent is targeted towards complex multi-agent systems and the estimated “how” type intent is represented explicitly in the system. The associate systems are integrated cognitive engines and all intent was represented in the graph structures, which are largely opaque to external AIAs and operators. Traces in the graph structure, referred to as a plan-goal graph by Geddes, attempt to model the entirety of intention, the “what”, “how”, and to an extent, even the “why” (Geddes, 1989). Operationalized Intent seeks to decompose the representation to make it more flexibly reusable by other AIAs in a multi-agent system through a focus on the “how” elements of intent with the perspective that the human operator will direct the “what” that is relevant to the AIAs and that the “why” is generally irrelevant to present AIAs.

Architecture and Ontology

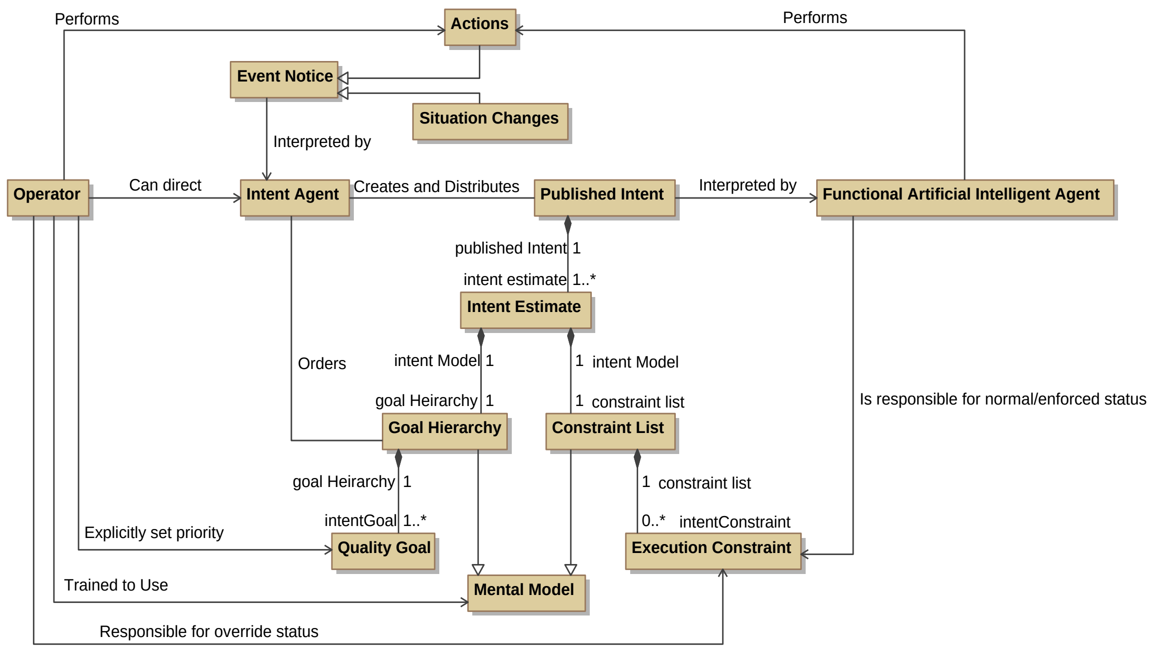


Figure 4. Ontology of Operationalized Intent

We have alluded to a conceptual architecture for Operationalized Intent which is worth summarizing to contextualize the discussion below. In an intent-informed system there are one or more human operators who interact with multiple task focused functional AIAs, and an intent agent for each operator. The intent agent is an AIA, the sole purpose of which is integrating intent estimates into published intent to be disseminated to the functional AIAs. There should be interfaces between the operators and the intent agent to support explicit coordination, and between the functional AIAs and the intent agent to support situation understanding by the intent agent and published intent dissemination to the functional AIAs.

An ontology of Operationalized Intent provides greater detail. Figure 4 depicts the definition of the relationships as a Systems Modeling Language (SysML) Block Definition Diagram. Since Operationalized Intent operates on the coordination cycle between the operator (on the left) and the functional AIAs (on the right) those blocks bound the discussion. Working from the bottom up, the quality goal and execution constraint form the basis of the published intent. The goal hierarchy and constraint list serve as the mental models that are common between the AIAs and the operator. These are managed by the intent agent whose sole purpose is to observe the available situational and operator data to produce intent estimates. The responsible functional AIAs determine the normal and enforced constraint statuses while the intent agent orders the goal hierarchy. An operator should always have the ability to direct changes to the AIAs internal state. Therefore they have the responsibility to set the override status of a constraint and the option to explicitly specify the priority of any goal in the hierarchy. The functional AIA and operator actions, along with any changes to the situation, both external and internal to the system, are the inputs to the intent agent and represented as event notices. These event notices are drawn from the Core Situation Awareness Ontology (Matheus, Kokar, & Baclawski, 2003) representation of situation evolution in time. The association between the operator and the intent agent is meant to clarify that the operator has access to direct the intent agent. However, to do so requires some operator action and would be an event notice. There are two “Interpreted by” associations, from the event notices to the intent agent, and from the published intent to the functional AIAs. These are the decoupling mechanisms that allow new information to be integrated into the intent estimation process independent of the functional AIAs (i.e. intent estimation),

while tailoring or tuning the use of the published intent by the functional AIAs (i.e. intent application). This allows the development of the functional AIAs by disparate groups with a common understanding of published intent. Application of this ontology to any given domain necessarily requires analysis of that domain to define useful quality goals and execution constraints, along with the relevant event notices and their interpretation. The Operationalized Intent domain application process is detailed in other manuscripts.

Conclusion

As AIAs in multi-agent systems become more capable of executing complex tasks, understanding intent is necessary to improve coordination of human-agent teams. We begin by reviewing the use of intent within the human factors and cognitive psychology literature which leads to the differentiation of intent estimation from intent application, as well as the differentiation of “why”, “what” and “how” based intent. A taxonomy of intent-based systems is then developed through a review of existing intent-based systems. Together these reviews show that intent has been modeled in a variety of ways without a cohesive understanding of intent and its different forms. Based upon these reviews and our understanding of multi-agent system architectures, we propose Operationalized Intent as a method of modeling intent regarding “how” the operators would like to execute the team’s tasks. We propose that by embedding knowledge of how to execute within a multi-agent systems, the available AIA may perform their tasks in a manner that is more useful and synchronized with the operators and other AIAs within the system.

IV. A Study of Operationalized Intent for Remotely Piloted Aircraft

Chapter Overview

This chapter provides the **Theorizing** phase discussion on methods to develop a quality goal hierarchy and study it. The **Studying** phase related to the quality goal rankings, intent change and cohesion are also included.

Effective teams coordinate their actions to achieve shared goals. In Human-Agent Teams, the Artificial Intelligent Agents (AIAs) struggle to coordinate effectively with human teammates as they lack an understanding of their human teammate's intent. As a result, the human teammate must explicitly communicate their task-oriented goals and how they are to be achieved. To improve the AIAs ability to coordinate, we have proposed a method to model situated operator priorities as a means to estimate "how" an operator desires a task to be performed, a construct we refer to as Operationalized Intent. Implicit in this construct is the assumption that trained operators will exhibit similar intent models within similar situations and that this intent will change with changes in situation. The focus of this paper is on the dynamics of intent and intent cohesiveness across operators. In this paper, we report the results of a study to track operator intent through a series of three tactical scenarios. The study employed an immersive, remotely piloted aircraft simulator to study intent in a synthetic task environment. Using operational pilots and sensor operators in realistic scenarios we were able to elicit their intent under relatively naturalistic conditions in the midst of challenging tactical situations. Analysis indicates that the Operationalized Intent method models intent which is dynamically responsive to changes in the situation and produces data that are suitably cohesive across operators to generalize to an operator role.

Introduction

Effective teamwork involves coordinated activity among teammates as they perform interdependent activities (M. Johnson et al., 2019). It is well recognized that humans estimate the intent of their teammates to anticipate and coordinate activity (Meltzoff, 2005). Past human-agent teaming research has proposed the estimation and application of intent to aid the coordination of artificial agents as they collaborate with their human counterparts (Riley, 1989; Rouse, Geddes, & Curry, 1987). Generally, these approaches have assumed that knowledge of the user's goals and the processes used to achieve each goal could be applied to anticipate human intent (Banks & Lizza, 1991). These systems have not gained wide acceptance. Lately, it has been proposed that it is important for human operators to communicate with Artificial Intelligent Agents (AIAs) regarding their intent to avoid miscommunication (Miller, 2017).

Early intent systems sought to understand and anticipate the goals that operators sought to fulfill. These systems decomposed functional goals to functional sub-goals at various levels of abstraction (Geddes, 1989, 1997; Rouse et al., 1987). When pursuing a goal at a lower level of this abstraction, the goal represents "what" the operator wishes to achieve while the higher-level goals represent "why" this goal is important. However, these frameworks also recognize that multiple means are available to achieve the desired goal and selection among these means requires knowledge of "how" the goal is to be achieved. It has been proposed that knowledge of "why", "what" and "how" tasks are to be achieved are important components within team mental models to facilitate effective teamwork (Rouse et al., 1992). Similarly, when attempting to design intent-based systems, one can determine "what" the operator wishes to achieve, as well as "how" they

wish to achieve the activity. Prior AIA modeling literature has also differentiated goals associated with the “what” and “how” differentiation. For example, Sterling and Taveteer discuss a framework for modeling human-agent interaction using functional goals as well as goals they term “quality goals” where it is recognized that “how” a functional goal should be achieved depends upon relative user priorities for items which are often mutually exclusive, such as performance, safety, and security (Schneider & Miller, 2018).

In the current research, we propose that “what” changes rapidly during execution. In contrast, “how” an operator performs the activity is sensitive to perceived operational constraints based on situational considerations. For example, fuel efficiency is likely critical for long-duration aircraft missions, but a minor consideration for short missions. While alternate goals, such as maximize maneuverability, might dominate a goal of maximize fuel efficiency for certain mission segments, maintaining high fuel efficiency will remain important throughout such a mission. Therefore, our working premise is that the priorities governing “how” an operator seeks to accomplish the activity is relatively stable with respect to the situation. This stability is important as it permits future human needs to be projected, enabling AIAs to anticipate future activity and plan coordinating activities.

In previous research, we have proposed a method of modeling operator intent related to “how”, referred to as Operationalized Intent (Schneider & Miller, 2018). By definition Operationalized Intent is comprised of a list of quality goals which are likely to be important to an operator within an operational context and their relative rankings. The ranking of quality goals provides information regarding relative importance of criteria and constraints which an AIA can use to inform task execution.

In this research, it is presumed that the operators reprioritize these quality goals in response to changes in perceived operational constraints and conditions. This motivates several research questions.

If Operationalized Intent is an internal mental model correlated to the situation, the study of intent requires a means of examining and comparing intent. Understanding internal mental states has been studied by cognitive engineering researchers through a variety of methods. This raises the question, what methods and considerations should be employed to elicit and understand user intent?

Does the operators' use of the quality goals indicate that it is effective in capturing intent? We posit that significant bias in the ranking of quality goals throughout a controlled, yet involved, tactical scenario may provide insight into their utility. Since the quality goal rankings are situationally correlated, quality goal ranking bias may also reveal strengths and weaknesses of the designed study and provide a means to refine it.

Based upon the prior discussion, we anticipate that the priority of quality goals should be stable over time unless a disturbance is introduced which induces a change in the perceived constraints. We expect the introduction of these disturbance to alter the user's priority of the quality goals. To test this premise, it is important to ask, how does a participant's quality goal priorities change over the course of a trial?

Finally, are the priority of quality goals cohesive enough across trained participants to permit them to be estimated using a common estimation model or algorithm? Specifically, we seek to assess if intent, when described by the priority of these quality goals, is cohesive across a group of highly trained and experienced individuals. This question is important since adequate consistency, as measured by the

cohesiveness of operators' goal rankings, enables agents to be designed and developed independently of the specific human operators in their team.

This paper presents the results and analysis of a study that tracks the priority of a set of quality goals across a set of experienced operators as they control a simulated Remotely Piloted Aircraft (RPA) through a series of three tactical scenarios. We begin by outlining the intent model development, study design, and metrics. We then discuss the study method, synthetic task environment, and execution. Finally, we present the analysis, which identifies changes in goal priority with changes in situation and the cohesiveness of the operators' intent models.

Intent Modeling

Background

To discuss intent we use Bratman's definition of intention: "relatively stable pro-attitudes that function as inputs to further practical reasoning in accordance with the two-level model of practical reasoning…" (Bratman, 1990). Researchers have modeled or operationalized this concept in multiple ways. In the Human Robot Interaction (HRI) and Human Machine Interface (HMI) domains, intent has been modeled as a physical path for prediction and error correction (Ahmad et al., 2016; Holtzen et al., 2016; McGhan et al., 2015; Periverzov & Ilieş, 2015; Vered et al., 2016). Others use a task or relational model to capture intent as a goal in an a priori defined set of goals (Bonchek-Dokow & Kaminka, 2014; Feng & Upenn, 2015). For example, the Associate System design represents intent as an implicit part of the acyclic directional graphs within their modeling technique (Banks & Lizza, 1991; Geddes, 1997). While initially designed to

estimate human intent without operator interaction, research involving the Rotocraft Pilot's Associate identified the ability of the human and the AIA to communicate on the current task and subtask as a key enabler to successful operation (Andes, 1997; Miller & Hannen, 1998). For our effort we sought an intent representation that is explicit, easily communicable, and information rich for both AIAs as well as operators. The intent model must be explicit to be shared among AIAs and understandable to the operator. To be useful in high stress and workload environments, it must be easy to communicate to avoid inducing higher workload during intent estimation failures. Finally, for an intent model to be useful in complex systems it must communicate a volume of information that provides context for mutual understanding between the team members.

Operationalized Intent Model Structure

We have proposed Operationalized Intent as a method of modeling operator intent in prior research (Schneider & Miller, 2018). It is targeted at trained operators utilizing complex systems in complicated environments to achieve a high degree of performance. The Operationalized Intent model is a shared structure for a mental model composed of two elements, an ordinal prioritized list of quality goals, and a list of execution constraints with their current status. The ranking of quality goals provides optimization guidance which an AIA can use to inform task execution. Effective quality goals abide by the following heuristics, they are: 1) relevant to the tasks the AIAs which are present in the system can perform, 2) conceptually relevant to the operators, 3) interpretable by AIAs to inform computation, 4) conceptually continuous, 5) violable by AIAs actions given other considerations, and 6) differentiable from other goals. Quality goals are

defined during design and only their relative priority (i.e. rank) changes during the dynamics of execution.

Model Development

Developing useful quality goals, which represent the mental model of an operator, requires an in-depth understanding of the cognitive work domain. The domain includes the system being employed, the environment in which it operates, and the work it performs. To capture and understand the domain we began with a Goal Directed Task Analysis (GDTA) (Bolstad, Riley, Jones, & Endsley, 2002) for the domain of interest and extended it to capture a quality goal model. While cognitive work analysis and work domain analysis (Naikar, 2017) are effective at capturing the work, environment, and systems to be studied, the GDTA's explicit modeling of functional goals provides a clear grounding for Operationalized Intent.

The GDTA begins by determining an overarching goal and decomposing it into a hierarchy of functional sub goals. Decisions associated with accomplishing the sub goals are then identified. The specific information requirements necessary to make the decision are then mapped to the decisions. This leads to a hierarchical, acyclic graph with the major goal being decomposed into sub-goals, each having an associated decision, which is linked to the individual information requirements necessary to support the decision. We employed the GDTA extensions of Humphrey and Adams, adding a nominal sequencing in the goal decomposition as a reference for Subject Matter Experts (SMEs) (Humphrey & Adams, 2011). Humphrey and Adams also categorized the information requirements into tools and resources, thought processes, people and groups, and SA information. To extend the GDTA for Operationalized Intent, we mapped quality goals

directly to the functional GDTA goals in the decomposition and mapped data elements to information requirements. Relevant quality goals provide prioritization regarding “how” the functional goal is to be achieved. Not all quality goals are relevant to each functional goal in the decomposition, e.g. quality goals about searching are not related to functional goals focused on attack. By establishing these relationships, the quality goals are directly tied to the domain and therefore are relevant to the operators and are observationally differentiable from each other. The data elements capture specific data available in the system which provide some relevant portion of an information requirement necessary to support an operator decision. This ensures that the quality goal hierarchy can be estimated situationally. Figure 5 provides the meta-model of this extended GDTA as a Systems Modeling Language (SysML) Block Definition Diagram. The core GDTA elements are in white. The Operationalized Intent specific extensions are in red. The yellow elements of the meta-model will be discussed in subsequent sections on trial design.

Source material for developing an intent model, like all GDTAs, can be gleaned from documentation, operator SME interviews, and observation of the work. In addition, the heuristics for useful quality goals and the available data elements require examination of the system design to understand the trades among priorities that the AIAs are capable of supporting. Interviews and discussions with the system design SMEs and review of design documentation are key to developing useful quality goals that can be estimated from the data elements and affect system operation. Once the intent model is established a proper study can be designed to evaluate it.

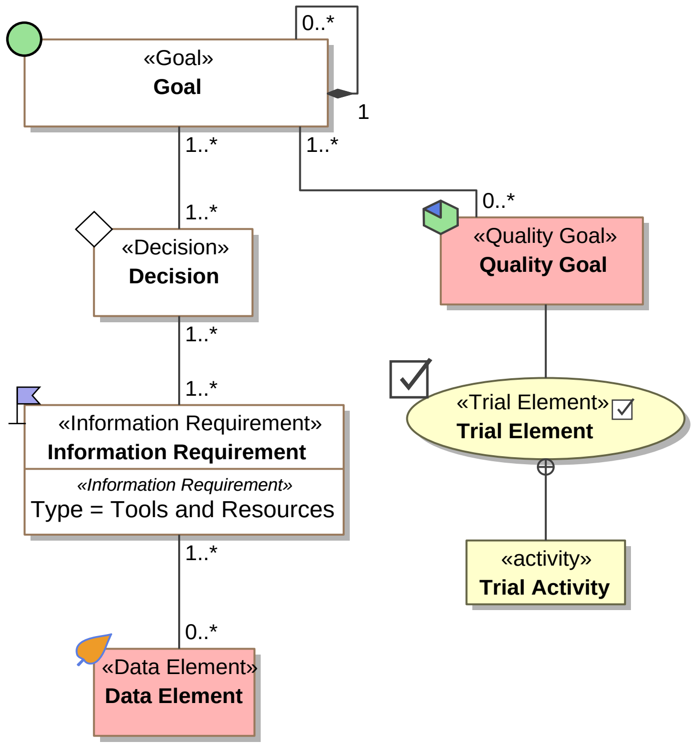


Figure 5. Meta Model of Operationalized Intent Goal Directed Task Analysis

Studying Intent

As a shared mental model, Operationalized Intent, is most appropriately examined in as naturalistic an environment as possible. When trained operators are situated in realistic conditions, working on tasks they perform regularly, the mental models they use can be examined with a high degree of validity (Endsley, 2000; Klein, 2008). Conversely, to study intent and capture the data elements for estimation, experimental control of the operational flow is necessary to ensure that changes in situation drive changes in intent. Balancing these competing priorities leads to studying operators in synthetic task environments (STEs) executing realistic operations.

Situating Operators

Properly situating operators to study intent requires commonality in their mental models, an environment of sufficient complexity to make the mental models necessary, and familiar work such that they use their mental models instead of developing new ones. Mental model commonality is the result of training and shared experiences (Entin & Serfaty, 1999). Selecting operators who have worked in the same or similar organizations with the same or similar systems based on similar training increases the likelihood of mental model commonality. The system and test environment complexity should be similar to the operational environment. Therefore ensuring that the sensory channels and channel loads (e.g. voice communication, visual scanning space, etc.), and that the complexity of information and control (e.g. spatial reference, text communication, detailed subsystem control, etc.) is similar in kind, if not in volume, to the operational environment. Finally, drawing on vignettes from actual operations that challenge the operators ensures that the work is familiar and motivates mental model use.

Trial Design

We propose that studying intent requires a synthesis of methods to inform the assembly of vignettes into trials. To examine Operationalized Intent during tactical operations we combined a dynamic situation awareness evaluation technique (Jamie C. Gorman, Cooke, Pederson, & Connor, 2005) with a static elicitation technique (Endsley, 2000). The trials were designed to be a series of stable situations interrupted by disturbances which would perturb the operator's intent before stabilizing into the next situation. This process of disturbing a stable state and observing the dynamic effects on situation awareness is based on the Coordinated Awareness of Situations in Teams

(CAST) methodology (Jamie C. Gorman et al., 2005). However, an operator’s internal mental model of intent is unobservable and, as yet, there is no basis for evaluating communication or interaction patterns to evaluate the operator’s intent. As such, the CAST method is not sufficiently detailed to study intent at the necessary depth. Since we need to elicit the operator’s internal mental model of intent at situationally representative points, the methodology design involved the experimenter pausing the simulation and requesting the operator rank the quality goals according to their immediate situation. This portion of the procedure was motivated by the Situation Awareness Global Assessment Technique (SAGAT) (Endsley, 2000). However, the current method did not require blanking the displays because there is no external representation of the operator’s internal intent available to the operator. For this study we assembled elements from vignettes based on RPA mission briefs, operational test scenarios, and training simulations. Using the meta-model in Figure 5, we mapped the trial elements to the quality goals to ensure that the disturbances are most likely to shift intent. To further improve the commonality of the operator’s mental models, we decomposed each trial element into activities and made sure that each activity was addressed during training so that the operators had a common understanding of the STE. To effectively study intent, the dynamics of CAST are combined with the detail of SAGAT and the context of naturalistic decision making study considerations.

Comparison Metric

The quality goal ranking elicitations during the trials provide insight to the intent mental model of the operator. To evaluate changes and similarity in intent models we must be able to compare those models using a computational metric. In studying intent,

we assume the quality goal hierarchy ranking changes from situation to situation and from operator to operator. To assess the changes in these ranks, we need to measure the difference between two sets of rankings. Many metrics have been used to compare ranked sets, including the Spearman Rank Correlation Coefficient (rho) and the Kendall Rank Correlation Coefficient (tau) (Vembu & Gärtner, 2011). The former uses the square of the difference in rank which penalizes large differences non-linearly. The latter is based on concordant pairs and measures distance based on the number of pairwise inversions to get from one set of ranks to the other (Dwork, Kumar, Naor, & Sivakumar, 2001). At this time there is no evidence that weighting the difference in ranking non-linearly is appropriate. Since the ranking of the quality goal hierarchy is independent from one situation or disturbance to the next, pairwise inversions are not relevant. The Spearman Foot Rule provides a distance between any two ranked sets of the same elements without any weighting (Diaconis & Graham, 1977). Since we are addressing goal hierarchies of less than 10 items, an exact, normalized version of the Spearman Foot Rule is used to improve accuracy. This modified metric is called divergence and is computed according to (1).

$$Divergence = \frac{2\sum_{i=1}^{n}|R_{A,i} - R_{B,i}|}{n^2 - (n \bmod 2)} \qquad (1)$$

where n is the number of quality goals in the hierarchy, $R_{A,i}$ is the rank of the i^{th} goal in the first set, and $R_{B,i}$ is the rank of the i^{th} goal in the second set. The results range from 0 to 1 in discrete increments, with 0 being perfect agreement and 1 being complete disagreement.

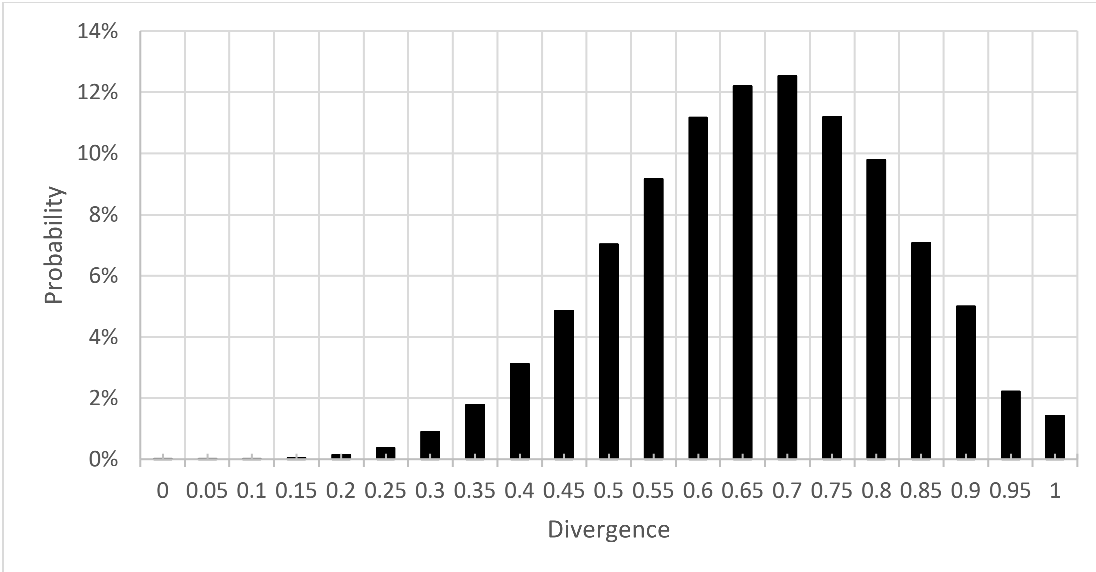


Figure 6. Probability Mass Function of Nine Goal Model Divergence

The current study employed nine quality goals, resulting in a metric having 21 even steps across the range which are spaced in 0.05 increments. A single step represents two adjacent goals swapping ranks (e.g. Goal A ranked 1 and Goal B ranked 2 in the first set and Goal B ranked 1 and Goal A ranked 2 in the second set with all other goals being ranked identically between the sets). There are 9! unique permutations of the quality goal model. The exact distribution of divergence between these permutations and any single permutation is approximately $N\left(\frac{2}{3}, \frac{8}{405}\right)$, the probability mass function (pmf) is given in Figure 6. This distribution provides a concrete baseline on which to assess the cohesiveness of intent elicitations. The cumulative probability of two independent rankings of the quality goal model being within the bottom half of the divergence range is less than 18.5%. If the observed divergence is distributed towards the lower half of the divergence range it indicates that there is some external effect driving the intent models closer together.

RPA Intent Study

Together, a rigorously defined intent model, a carefully designed study, and a comparison metric permit the study of intent changes over the course of a tactical operation. This section details the specifics of a study conducted with RPA operators.

RPA Intent Model

Following Endsley's process for developing GDTAs, we began by reviewing documentation on how the RPA is employed, the operational environment, and the capabilities of the simulation system for the study. This system contained only two AIAs, the autopilot and an Advisory-Cautions-Warnings system, so we extended it to a future concept that included additional AIAs to assist the operator, including an auto-router, sensor tasking management, defensive systems, weaponeering, etc. These AIAs opened the trade space to allow more goals to be relevant to AIA tasks. Coupled to the document review we performed multiple subject matter expert interviews with current RPA pilots to inform and validate the GDTA and the quality goal model. The result was a 9 quality goal model across the two most common missions in the RPA community. The quality goals are listed in the elicitation interface, Figure 7, which requires the operator to drag each quality goal from the left box to the right box. There are natural tensions between these quality goals. Maximize Time on Station means that fuel efficiency should be maximized to ensure a long time aloft. This is in tension with Time on Target and Maximize Maneuverability which both trade fuel efficiency for speed, either to get to the target on time or to maneuver dynamically at altitude. Due to the diversity of the missions flown by RPAs, not every quality goal is relevant to each situation. Maximizing Search Efficacy is not relevant when attacking a target. Conversely, Maximize Reattack

Opportunity is only relevant when the target is likely to be attacked. Both of these impact altitude, flight path relative to the target, potential terrain masking, bank angle limits, subsystem controls, etc.

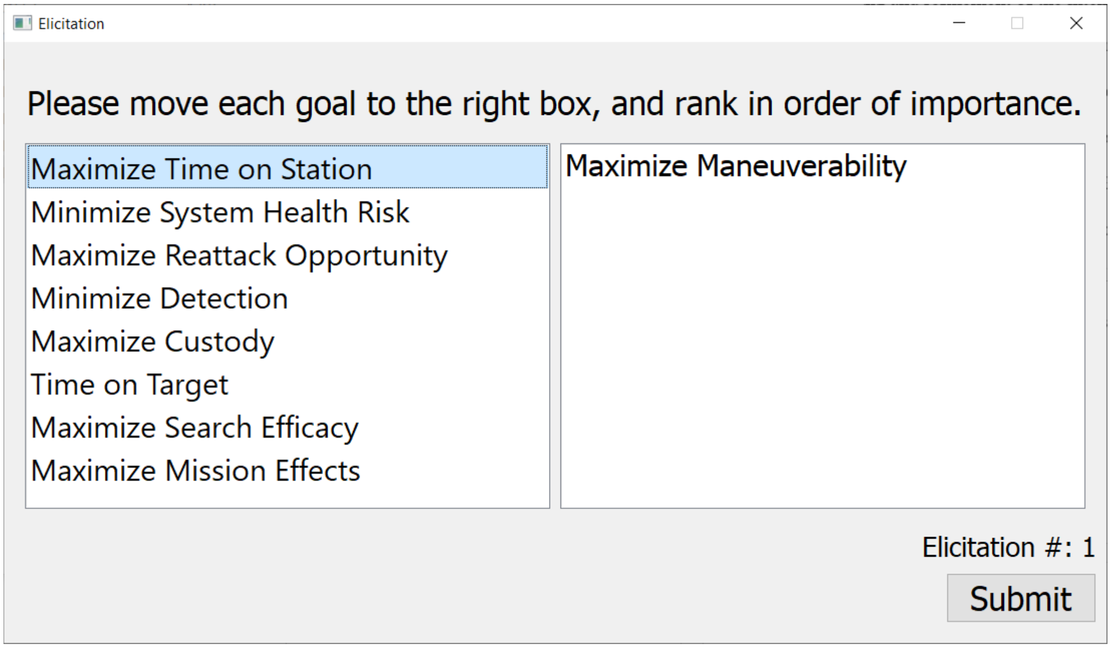


Figure 7. Quality Goal Model in Elicitation Interface

Operators

Our operators were drawn from pilots and sensor operators in an operational RPA squadron. All seven operators (four pilots and three sensor operators) were male having a mean experience of 1489 flight hours (st. dev. 558 flight hours) in RPAs. They all routinely performed the ISR mission. Each of the four pilots indicated that they perform CAS routinely, while the three sensor operators indicated they perform CAS infrequently. Each participant was randomly assigned a two digit identifier.

Synthetic Task Environment

The operators were experienced in flying their operational ground control station with a pilot in command (PIC), sensor operator (SO), and a mission intelligence

coordinator (MIC). Their current operational system is highly federated, requiring the operators to mentally fuse data from multiple sources and provide inputs to multiple systems using different means with dedicated controls at separate stations. The experimental STE used in this study is a fully integrated control station which allows a single operator to control the entire system. This STE provided a substantial shift from the operator's current environment, so for the study an experimenter performed the role of the MIC which is to back up the operator on video surveillance, assist with communication, and research historical and tactical information relevant to the situation.

The operators were seated at the STE crew station composed of two vertically arranged, horizontally oriented, 43 inch 4K displays. The upper display was mounted perpendicular to the floor, the lower display was reclined away from the operator at a 35 degree angle from the floor. The operator had a mouse, keyboard, and a noise cancelling headset with microphone. To avoid glare on the reclined display, all overhead lights were covered. Indirect lighting on the wall behind the workstation and the ceiling provided adequate illumination for the operator. The operators were seated in an adjustable office chair. The experimenters were seated in the same room, behind the STE crew station at another dual display computer workstation with headset and microphone. This enabled simulated live voice communications between the operators and other simulated mission players. The study was conducted between or before shifts in a squadron relaxation room at the Michigan Air National Guard base in Battle Creek, Michigan.

The STE crew station ran the Air Force Research Laboratory Vigilant Spirit Control Station software which provides a modular, tool based layout to control all aspects of the RPA (Rowe, Liggett, & Davis, 2009). This included chat, voice

communication, full motion video, alerts, and checklists on the upper display, with subsystem controls, navigation, and an integrated tactical map on the lower display. To rank the quality goals the operators used a simple drag and drop graphical user interface (GUI). The simulation was controlled from the experimenter station through the Vigilant Spirit Simulator. The experimenters also had access to a control station instance to observe the operator's view of the situation and serve as the MIC.

The missions flown in the study represented prototypical missions in recent operations. There were multiple external mission players, including customers, air traffic control, and ground forces, which the operator interacted with via chat or voice communication. The surveillance targets and mission taskings were similar to real world experiences. The simulated world was approximately 4 km square. The three trials took place in different sections of this area to provide a sense of uniqueness for each trial. This environment was designed to provide an immersive environment to permit the operators to achieve a tactical mindset from which we can elicit realistic rankings of the quality goals.

Trials

From the GDTA and intent model we drew upon operational test scenarios and published accounts of missions which resulted in meritorious awards for the crew. The result was three trials containing at least three disturbances with 5 to 7 elicitations taken at situationally relevant intervals. Each trial began with a different form of ISR mission and progressed to a complex CAS mission before being resolved.

Trial 1 consisted of a "Pattern of Life" mission (observing a compound for activity). Situation 1 and 2 each involved observing the compound for activity. These

situations were interrupted by an airspace restriction which threatened the current route and the emergency route home, disturbance 1. During the scenario a “High Value Individual” leaves the compound in a car, disturbance 2, and makes several stops, resulting in situation 3. At the final stop the individuals engage in hostile activity and the operators receive an order to strike the individuals, disturbance 3. They successfully carry out the strike resulting in situation 4.

Trial 2 begins with an area search for targets, which is interrupted by an aircraft generator failure. The operator is able to reset the generator and resume searching for targets. Once they have located and identified the targets and confirmed the most critical target they are given orders to strike that target. They then shift to supporting another manned aircraft strike on one of the targets the operator located. That manned aircraft is lost and they immediately initiate personnel recovery. The trial ends when they locate the pilot.

Trial 3 opens with the aircraft far away from the operational area, on the wrong side of a thunderstorm, with a fixed time on target required to rendezvous with their supported unit. With precise navigation and increased air speed they can just make the rendezvous in time. The mission is to support a convoy and identify obstructions or risks. Once the operator identifies the obstruction to the convoy route, the convoy re-routes and proceeds on their way. During the over watch, the aircraft link to the ground station is lost and the operator has no control or information from the aircraft for a short period of time. Once link it restored the operator follows the convoy to a crowded compound. An adversary convoy rolls in and a firefight commences. The operator is tasked with striking the adversary convoy to break contact and allow the friendly forces to regroup. The strike

destroys the target and two other adversary trucks flee. The operator must track them and mark their locations to end the trial.

These trials are highly dynamic with many shifting priorities, disturbances, and time sensitive situations. Figure 8 provides a comparison of the simulated timeline for each operator in Trial 1. The dark grey segments indicate disturbances, the diamonds mark when elicitations were taken. Some operators react more quickly to disturbances, some take their time with certain tasks to ensure they are complete, and some get lucky and achieve their tasks rapidly. We elicited their intent at situationally defined points during the mission which is why the timelines across participants do not line up. The elicitations, however, are situationally similar (e.g. all the "1" elicitations occur during the first situation).

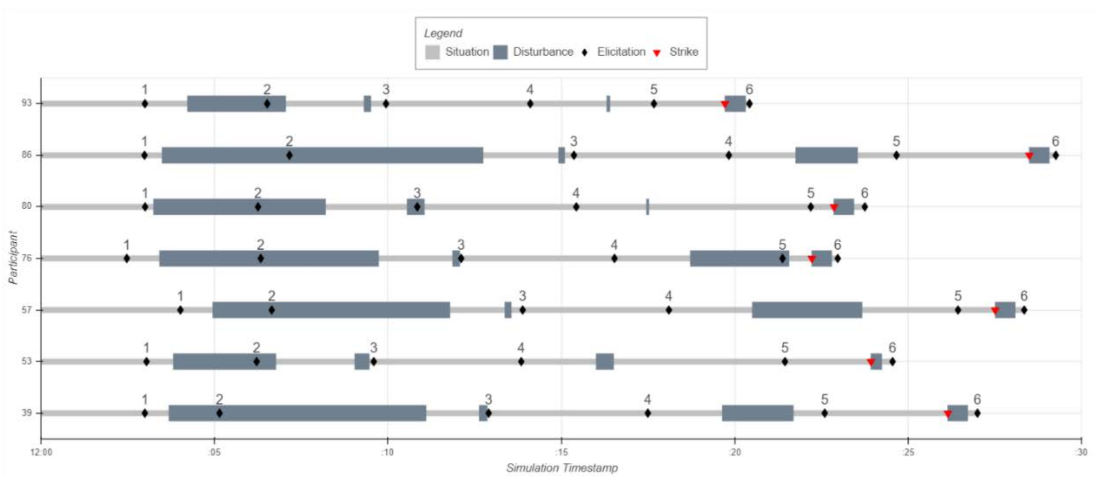


Figure 8. Simulation Timeline of Trial 1 by Operator Number

Procedure

The operators were provided with an informed consent document and took an initial demographic survey. They were then briefed on the study, including the quality goal model, the synthetic task environment, and finally the missions and trial flow. It was

stressed that the study was designed to understand tactical thinking and mental models so mission performance was not being assessed. Operators were instructed that not all quality goals were relevant in all situations and during an elicitation to focus on prioritizing the relevant quality goals and to place the irrelevant goals in any order at the bottom of the list.

Due to the significant differences between the current ground control station and the Vigilant Spirit control station, detailed training was necessary to ensure the operators could adequately perform the trials. The training took about an hour and included in-situ walk-throughs of the elicitation GUI and the entire Vigilant Spirit control station. This included all tasks, tools, and event types the operators would encounter during the trials. During this time operators could ask any questions and revisit any portion of the training to better understand the flow. Since the focus of this study was on tactical thinking, not mission performance or interface design, the operators were informed that if there was any time they felt lost in the interface or were unsure how to access a needed control, the simulation could be paused to clear the confusion.

Trials were presented in a blocked sequence so that order effects would be controlled. They were scheduled for an hour, but none lasted more than 40 minutes of simulator time. At the conclusion of a trial each operator filled out a short survey on their intent change. After each trial the operator was afforded a break while the next simulation was setup. Some operators performed all three trials over the course of a single day, with most conducting the trials over a few days, and a few over the course of a week. All data was collected within two weeks. At the conclusion of the study, each operator filled out a

final survey which focused on the overall concept of intent informed Human-Agent Teaming.

Data Collection and Analysis

A PostgreSQL database logged all the data from the elicitations and situational data from the control station. No mission performance data were collected during the trial. All operators were able to successfully accomplish each mission, with one exception in Trial 2 that was due to a failure of the simulator.

This paper focuses strictly on the elicitation data, the situational data will be discussed in a future paper. The elicitation data were stored as a table with each elicitation representing a row which contained the ranks for a given quality goal column. The data were cleaned by eliminating the elicitations during training sessions and operators who did not complete the entire study, and any redundant elicitations. To ensure that the elicitations are situationally similar, their timestamps were compared to the situated events which denote the disturbances and situations. Figure 8 depicts the sequence of elicitations, situations, and disturbances based on simulation timestamp for Trial 1 for each operator. The first elicitation comes during the first situation. The second elicitation was taken during the first disturbance. The third elicitation was taken in close proximity, during or immediately after, the second disturbance. The fourth elicitation is during the third situation while the fifth elicitation follows the third disturbance but occurs before the strike. The sixth elicitation was taken after the strike was completed. While some situations and elicitations occur at similar times, as the trial progresses the time necessary for each operator to reach the situation for a disturbance or elicitation

varies. Each elicitation is situated in similar mission phases (i.e. during the same situation or adjacent to the same disturbance) for the purpose of analyzing across operators.

The analysis for these data focuses on the previously stated research questions. Is the intent model effective in capturing operator intent? Distributional observation of the ranks assigned to each quality goal, across trials and operators, is used to identify biases that can inform intent model improvements. Next, does an operator's intent change over the course of a trial? If intent never changes then it is unnecessary to estimate it. To evaluate this question, comparisons are made between sequential elicitations as well as by comparing the maximum divergence between elicitations within each trial. In addition, the distribution of divergence between all elicitations for an operator within a trial are evaluated. Finally, is the intent across operators cohesive enough to indicate it can be estimated using a common algorithm? If the intent of similar operators in similar situations varies widely it may indicate that generalizing intent to a role is unrealistic. We ensure situational similarity by coding each elicitation according to its situation and then comparing them within a situation and across operators. The simulator failure of operator 93 in trial 2 resulted in their data being removed from the cohesion analysis because it could not be consistently situationally coded.

Table 9. Heuristic Threshold Examples

Quality Goal	Minimum Threshold (0.1)		Top/Middle Threshold (0.2)		Top/Bottom Threshold (0.4)	
	Rank A	Rank B	Rank A	Rank B	Rank A	Rank B
A	1	2	1	5	1	9
B	2	1	2	2	2	2
C	3	4	3	3	3	3
D	4	3	4	4	4	4
E	5	5	5	1	5	5
F	6	6	6	6	6	6
G	7	7	7	7	7	7
H	8	8	8	8	8	8
I	9	9	9	9	9	1

To address these qualitative questions, we established several heuristic thresholds laid out in Table 9. The minimum threshold is a change in rank of at least 4 goals, a divergence of 0.1. The cumulative random chance probability, the summation of Figure 6, of observing divergence at 0.1 or below is 0.014%. The inversion of any two adjacent goals is assumed to be of negligible consequence, therefore we set the minimum threshold at two sets of adjacent goals. The top/middle threshold was computed by swapping the ranking of the first and fifth ranked goals, a divergence of 0.2 at a cumulative probability of 0.204%. Given that not all quality goals are relevant to every situation, it is possible that the bottom set of goals are irrelevant to the operator. This threshold is therefore the difference between the top goal being the same or being potentially considered irrelevant. Finally, the top/bottom threshold was computed by swapping the ranking of the first and ninth ranked goals, a divergence of 0.4 at a cumulative probability of 6.372%. This represents a significant difference of perspective on the situation. Note that it is possible to have divergence at or higher than these thresholds and still have the top rankings the same, these are the minimum thresholds

below which it is impossible for these conditions to occur. These thresholds represent the low divergence tail of the random chance distribution which indicates that if the research questions achieve these thresholds they are unlikely to be random. While the divergence metric is approximately normally distributed, there is no expectation of normalcy in the data and the sample sizes are small, therefore nonparametric tests are employed in the analysis.

Results and Analysis

Trial and Intent Model Evaluation

Initially, we assessed the utility of the quality goals based on the distribution of their rankings. All quality goals were assigned the entire range of ranks, with the exception of “Maximize Custody” which is not assigned a rank of 9. Figure 9 plots the quality goals versus the rank, colored by participant with the symbols corresponding to the trial. The data are jittered to provide a sense of the distribution. In addition, the median and interquartile range (IQR) are plotted for reference. Three quality goals (Maximize Custody, Maximize Maneuverability, and Maximize Mission Effects) have medians above the median of the rank range (4.5) indicating that they are always ranked highly. Two quality goals (Maximize Time on Station and Time on Target) have IQRs below the median of the rank range which indicates they were seldom rated highly. In fact, Maximize Time on Station is only ranked as the highest priority once.

Overall these results indicate that the intent model captures the range of intent and the trials presented situations that varied intent as measured through the priorities of the quality goals. Due to the dynamic and active nature of the trials, the operators are almost

always acquiring and maintaining custody of a target. This is core to the RPA mission so it is understood that "Maximize Custody" would be rated so highly. The trials were short and the fuel always started full which inherently places less emphasis on the "Time on Station" goal. In addition, because time was relatively short, only Trial 3 had a "Time on Target" situation and it was only truly relevant for one or two elicitations which explains its low rank distribution. The results indicate that "Maximize Mission Effects" does not effectively differentiate situations because it has high median and a tight IQR which indicates little change over the course of the trials.

Figure 9. Jitter Plot of Quality Goal Ranks, Trials are designated by shape, Participants are designated by color.

Intent Change

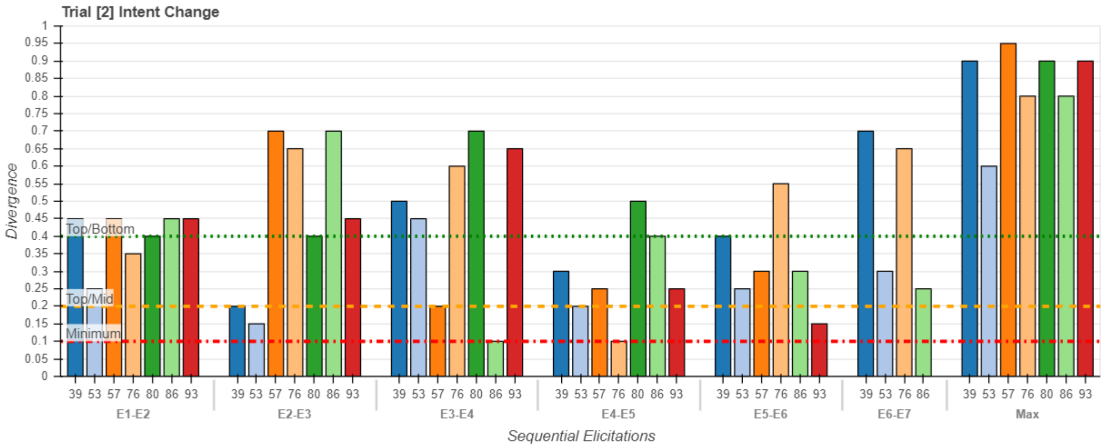


Figure 10. Trial 2 Divergence between Sequential Elicitations, colored by operator number.

For intent change, higher divergence with changes in situation indicates greater change in intent. Figure 10 provides the divergence of the operators ranking for each sequential pair of elicitations as well as the maximum divergence between elicitations for Trial 2. The divergence for each operator is coded with a different color and sequential elicitation pairs group the operator numerical designations. Across all trials, the sequential elicitations, within themselves or across operators do not exhibit any consistent pattern. Each operator has some divergence values above the minimum and the top/middle thresholds and some divergence values at or below the minimum threshold. The final grouping shown in Figure 10 represents the divergence between the ranks from the two elicitations with the maximum divergence. These results indicate that while the individual steps from one quality goal hierarchy to the next varied in their divergence as a function of change in situation, the maximum divergence between any two elicitations in the trial was above the top/bottom threshold. Across all operators and trials, the maximum divergence is significantly above the threshold based on a Kruskal-Wallis test (n=21,

p<0.05). The distribution of pairwise divergence of each elicitation with the others, for an operator, within a trial is given in the boxplots of Figure 11, the whiskers denote the range. A Kruskal-Wallis test found that the median divergence of all operators was significantly above the minimum and top/middle thresholds (n=15, $p < 0.05$) for all trials. For the top/bottom threshold, across all trials, twelve were significantly above, one was significantly below, and eight were not significantly different. There were no instances of zero divergence, i.e. no two elicitations provided identical responses.

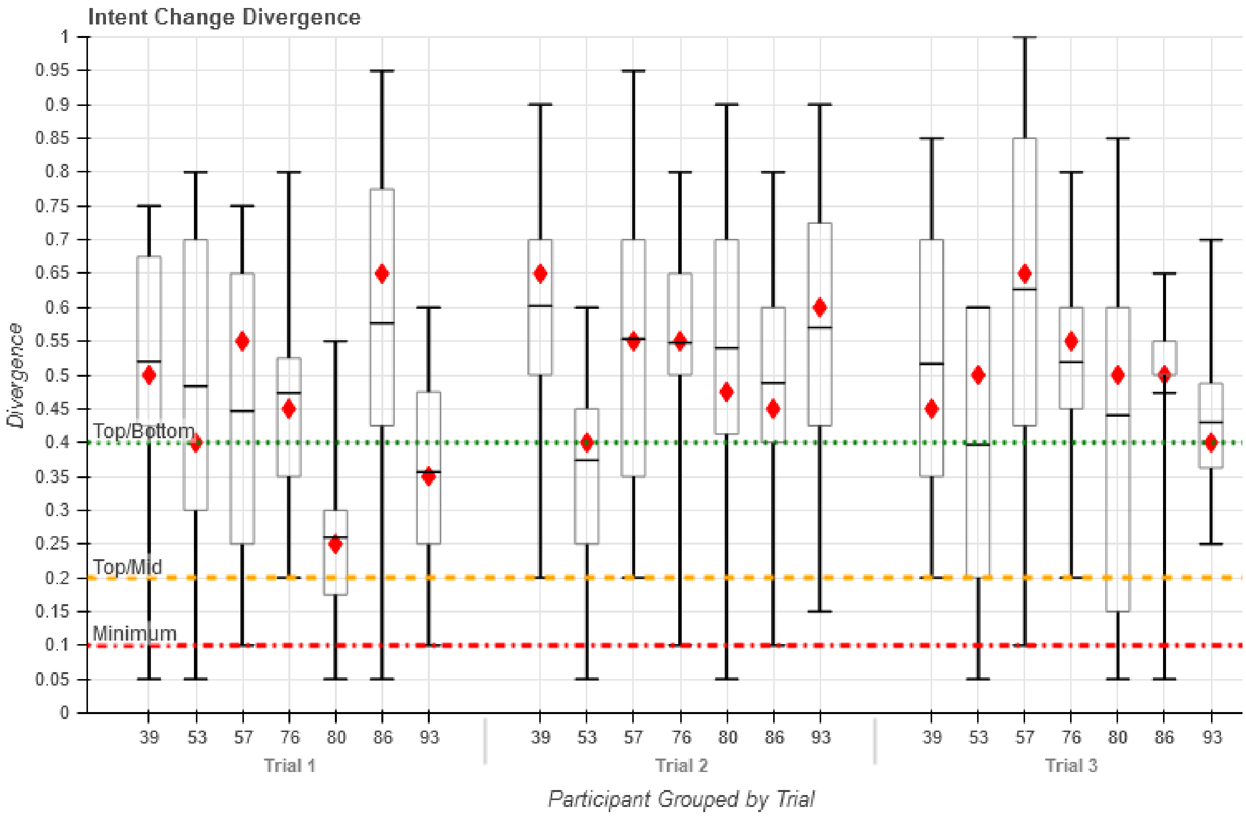


Figure 11. Participant Intent Change Divergence Grouped by Trial. Bars represent the range and the heuristic thresholds are overlaid for reference.

These results indicate that intent is changing during the trial. While the sequential elicitations do not consistently exceed the minimum threshold, these values do consistently exceed the top/bottom and minimum threshold values. The change in intent

over the course of the trial is above the top/bottom threshold and the median intent change is above the top/middle threshold.

Intent Cohesion

In this cohesion analysis, the desired outcome is for divergence to be below a threshold when comparing across operators. This result would indicate that each of the operators have a similar intent at each situation within the mission. Figure 12 depicts the divergence matrix and distribution of divergence values for Elicitation 4 in Trial 2. Each elicitation is designated by a code: Operator number, Trial number, and Elicitation number (e.g. P76T2E4 is operator 76, trial 2, and elicitation 4). The histogram provides a distributional view of the matrix data binned to the discrete divergence steps. The top/bottom heuristic threshold is provided as a dotted line. The dashed line is the probability mass function of the exact discrete distribution of the divergence for a 9 quality goal model. Using a Kruskal-Wallis test, all elicitations in all trials were significantly above the top/middle threshold ($n=21$, $p < 0.05$). For the top/bottom threshold, twelve elicitations were significantly above threshold, five elicitations were not significantly different from the threshold, and two were significantly below threshold ($n=19$, $\alpha=0.05$). The distribution of the divergence data for each elicitation was compared to the known exact distribution for a nine quality goal model using a χ^2 independence test. To ensure independence between tests, a random sample, equal in size to the divergence matrix, was drawn from the exact distribution and unequal bins were developed to ensure at least four bins, with at least four counts each, were present in the sample distribution. These bins were then used to build a histogram of the divergence data for comparison. The χ^2 test for each elicitation indicated that the operators' intent

divergence values were significantly different from random chance for all except two elicitations. These two elicitations contained data for only four participants and were not significantly different (n=19, α=0.05). These results indicate that while the operators were more cohesive than random chance, there was a 50% probability that the rankings provided by any two operators were as large as would be obtained if any operator could have their top ranked quality goal ranked by another operator as their bottom quality goal without changing any other rankings.

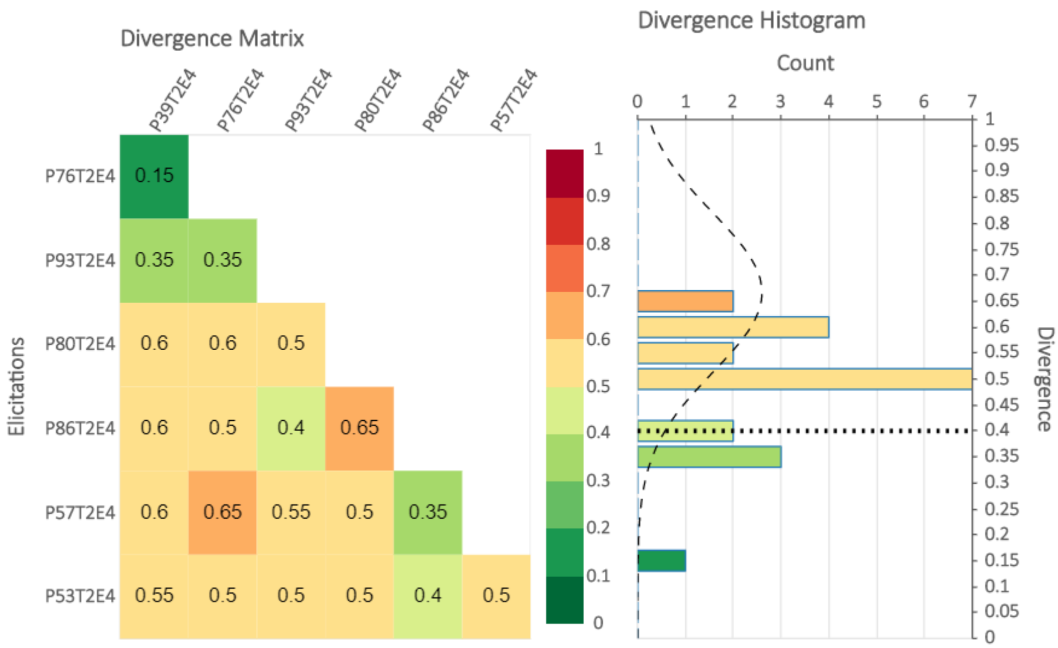


Figure 12. Example Divergence Matrix and Histogram. The dotted line in the histogram is the top/bottom threshold, the dashed line is the probability of randomly guessing.

For intent to be generalizable to a role requires that operators respond to similar situations with similar intent. For each elicitation we tested the distribution of pairwise divergence values for each operator against the other operators. In the Kruskal-Wallis test (n=6 pairwise divergence values per operator, m=7 operators) for the 19 elicitations across all trials, all but three failed to reject the null hypothesis (α=0.1). This means that

for three of the elicitations the median divergence of each operator had low or high divergence contributors. A low divergence contributor is one where the operator is very similar to all the other operators (i.e. close to the center of an evenly spaced group). A high divergence contributor is different from all other operators (i.e. outside a cluster). The Conover test was used to identify the operators that were significantly different. Each elicitation indicated that there were operators who provided significantly different rankings from at least 3 other operators (α=0.1). In some cases, these were low divergence contributors. For each trial we took the median divergence of each operator, and the median divergence of the elicitation, and used a normalized ranked sum to develop a contribution score for each operator to identify those most responsible for high divergence. Figure 13 indicates that, operators 93, 57, and 53 are responsible for more divergence than any other operator.

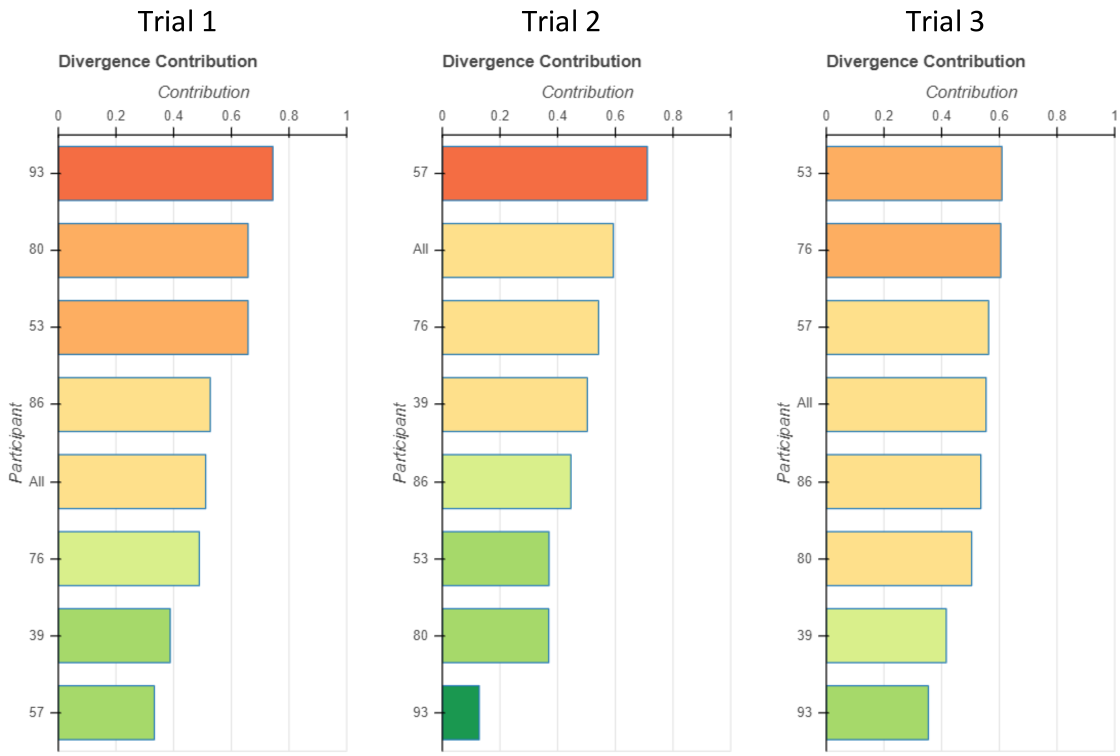


Figure 13. Divergence Contribution Scores for Each Trial

Omitting each top contributor, and participant 93 who experienced a simulator malfunction during trial 2, from the analysis yielded substantially better results. In 19 elicitations, three were significantly below the top/bottom threshold, twelve elicitations were not significantly different than the threshold, with four elicitations being significantly above the threshold ($n=15$, $\alpha=0.05$). The Kruskal-Wallis across operators failed to reject the null in all but the fourth elicitation of Trial 1 ($n=15$, $\alpha=0.1$). The contributor elicitation Conover matrix indicated that two operators were significantly different from the other four, but these two operators had the lowest median divergence, meaning they were centrally located, low divergence contributors. By omitting high divergence operators, the divergence distribution is not significantly above the

top/bottom threshold and the operators have similar divergence indicating that their intent is cohesive.

Discussion

The results indicate that Operationalized Intent was effective at tracking changes in operator intent through a tactically evolving situation. The training on Operationalized Intent was short and the operators were inexperienced in using it, they nonetheless demonstrated a reasonable level of cohesion, when compared to our heuristic threshold. This indicates that with further training and integration of Operationalized Intent into tactical instruction, generalizing a model to the role of an operator is likely feasible. Operationalized Intent also provides a means of identifying differences in tactical thinking between operators within these simulated environments. The high divergence contributors motivate tailoring intent estimation efforts to an individual since the cohesion remains broadly distributed between 0.9 and 0.1.

While individually, the sequential changes in intent, as represented by changes in goal priority, were not necessarily significant, over the course of a trial they represent a significant shift. This demonstrates that “how” type intent does not necessarily fluctuate significantly over a single change in situation, but evolves through multiple situation changes.

Assessment of the intent model indicated that one quality goal should be reassessed. “Maximize Mission Effects” seems to lack differentiability and may be more appropriately be re-designated “Maximize Weapon Effects.” In an iterative design process this kind of update could be made and rolled into further evaluations. Studying

Operationalized Intent in longer, more realistic missions may demonstrate what the actual distribution of ranks should be for a quality goal. This can be used to further refine the intent model.

To define and operationalize a model of intent demands cognitive engineering analysis, but to study it effectively requires multiple concurrent evaluation methods. By synthesizing dynamic and static situation awareness assessment techniques we developed a means of situating operators to enable the study of intent. Using the situation-disturbance framework, the trials presented the operators with a dynamic environment which demanded they shift intent to respond to the next challenge. In addition the method we present is scalable beyond the single operator or single system paradigm to a multi-operator team where the communication tracking of CAST and multiple perspective elicitation of intent through query techniques similar to those applied in SAGAT can illuminate role differences across the team. This could enable intent estimation for each team member individually so that the AIAs can be most useful to each operator and the team simultaneously.

Through extending the GDTA, we successfully designed an Operationalized Intent model that could capture the operator's quality goal priorities in a computational representation and identify how they evolved over a dynamic tactical situation. This explicit model can be shared with other AIAs and provides a domain specific vocabulary to expeditiously coordinate "how" activities are performed within a human-agent team. Multi-agent systems are the reality of the near future. Highly integrated associate systems (Geddes & Buchler, 2012) face significant sustainment challenges as the other agents in the system evolve and operations change. Understanding a larger context, distilled into a

computational form, as we have demonstrated with “how” type intent, allows narrow AIAs to focus on estimating and applying “what” type intent specific to their tasks. The ability to express “how” type intent through a specified vocabulary provides a method for structured communication between human operators and AIAs within a multi-AIA system which govern the performance of the human–agent team.

Conclusion

The current paper proposed a method for studying operator intent, targeting “how” a goal is to be achieved. This proposed method extended GDTA to develop a method for deriving quality goals associated with operator intent pertaining to “how” a goal is to be achieved. Additionally, the method borrowed from the situation awareness literature applying situation changes similar to CAST and an elicitation method similar to SAGAT, to permit changes in quality goals associated with changes in “how” intent to be studied within a complex remotely piloted aircraft synthetic task environment using experienced operators. Changes in quality goal ranks resulted from disturbances which were injected into the simulation which modified the situation. Further, reasonable agreement in quality goal ranking was observed between the experienced operators, although some operators displayed notable differences from the other operators within specific trials. Overall, it is believed that this method can provide a real time understanding of changes in quality goal priority associated with ”how” intent. We propose that this method may be useful in obtaining a model of operator “how” intent which can be employed to improve the intent based communication between AIAs and

operators in a human-agent team, improving team coordination to improve coordinated task execution.

V. Operationalized Intent Estimation Accuracy

Chapter Overview

This chapter encompasses the **Theorizing** phase discussion on situational data, the **Studying** of intent cohesion dynamics, and the **Estimating** phase of the research method.

Teams of human operators and artificial intelligent agents (AIAs) in multi-agent systems present a unique set of challenges to coordination among team members. This research endeavors to address issues of estimating operator intent to provide the AIAs with a representation of the operator's internal mental model. Using a representation referred to as the Operationalized Intent model to capture quality goals relevant to "how" the operator would like to execute the team's mission, this paper details the development and evaluation of a random forest estimator to estimate operator priorities. By employing cognitive engineering analysis an intent model relevant to the entire human-agent team is developed along with the situational data to estimate intent. Estimation is structured as a label ranking problem in which quality goals, which dictate "how" work is to be conducted are ranked according to their priority. Modifying an existing label ranking algorithm, we demonstrate that the Operationalized Intent Estimator – Random Forest (OIE-RF) can estimate intent more accurately than the situation baseline. OIE-RF demonstrates stability in dynamic testing and the ability to use explicit communication and knowledge of the operator to increase accuracy. This exploratory research opens a new avenue for improving coordination and performance of human-agent teams.

Introduction

With the inexorable increases in artificial intelligence performance and diversity, the challenge of integrating artificial intelligent agents (AIAs) with human operators to form a synergistic and effective human-agent team hinges on their ability to coordinate. Mutual observability has been demonstrated to improve coordination (Schaefer, Straub, Chen, Putney, & Evans, 2017). There is a need to predict future behavior to permit one team member to plan actions in coordination with another, as well as, to direct or negotiate changes in another's future action (M. Johnson, Bradshaw, & Feltovich, 2018). For this reason, AIAs which include at least rudimentary methods for understanding a human teammates' intent are becoming increasingly common (Banks & Lizza, 1991; Bonchek-Dokow & Kaminka, 2014; Geddes, 1989; Huang et al., 2017; Kelley et al., 2008; McGhan et al., 2015; Vered et al., 2016; Xing et al., 2019).

Understanding human intent is challenging for today's systems as current day AIAs are narrow in scope and need to not only estimate human intent but to determine appropriate coordinating activities in response to changes in human intent. This problem is compounded when multiple AIAs are teamed with operators as is common in complex systems. If intent estimation and application are performed by each AIA, the understanding of intent has the potential to be as diverse as the AIAs and the designers who build them. To address this problem we have proposed federating the estimation of intent from the application through a design pattern called Operationalized Intent (Schneider & Miller, 2018). The concept involves the design of a single agent whose sole purpose is to form a cohesive estimation of an operator's intent. This intent estimate is then provided explicitly to other team members. This estimate then provides execution

guidance to other AIAs to supplement the necessary explicit communication within the team.

Operationalized Intent

The formulation of Operationalized Intent differentiates "what" from "how' based knowledge as has been recognized in research on human mental models (Rouse et al., 1992). Further, it accepts the premise that explicit communication is likely necessary to specify "what" must be performed, particularly as this communication pertains to more abstract concepts (Miller, 2017). However, the formulation proposes that "how" based knowledge is important to communicate and situationally-driven, making it more stable with time (Schneider, Miller, Jacques, Peterson, & Ford, 2020). Further, we have proposed that the model of intent can be explicitly communicated with the human operator through an ordinal prioritized ranking of quality goals and a list of execution constraint statuses (Schneider, Miller, & McGuirl, 2020). Quality goals are execution relevant guidance that expresses how tasks are to be performed, not which tasks to perform. For example, "Maximize Fuel Efficiency" or "Minimize System Health Risk" doesn't inform an AIA on what to do, but in a prioritized list they provide a model of how important it is to achieve desirable fuel consumption versus safe operation. Thus, the prioritized list of goals permits the real time tradeoffs in system parameters when executing activities with a moderate or high level of abstraction. To be useful quality goals should abide by the following heuristics: 1) relevant to the tasks AIAs can perform, 2) conceptually relevant to the operators, 3) interpretable by AIAs to employ computationally, 4) conceptually continuous, 5) violable by AIAs actions given other

considerations, and 6) differentiable from other goals. Quality goals are defined during design and only the ranking within the hierarchy changes during execution.

The prioritized quality goal list operates as a shared mental model between operators and AIAs. As such, the prioritized list requires careful development to ensure it is applied consistently to remain shared knowledge between the human operators and the multi-agent AIA system. The purpose of this research is to explore development of a method to estimate relative goal priority. Specifically, we are interested in the accuracy of the estimates based on their distance from the elicited ranking. An estimator whose estimates vary dramatically during a stable situation would require constant correction by the operator to be useful. So, the stability of the changing estimates is of interest. Conversely, an estimator which does not respond to shifts in situation is unhelpful. The situational sensitivity and ability of the estimator to refine the estimate over time are two key performance attributes of interest. Finally, research indicates that explicit communication between team members and tailoring responses to the specific individuals improves estimates. Therefore, multiple estimation models with varied feature sets examine the effects of communication and personalization.

This paper begins by exploring a method for developing the intent model and to identify the data useful for intent estimation, as well as the conversion of the data into a computational representation. We then detail a potential label ranking estimation algorithm based on Label Ranking Random Forest (LRRF) (Zhou & Qiu, 2018).We then discuss the performance results of both static and dynamic tests in which the resulting model's predictions are compared to operator priorities.   Finally, we conclude by

discussing the utility of this method for providing a robust estimate of an operator's quality goal priorities.

Data Model Development and Acquisition

The proposed intent model is an explicit mental model of the operator's internal priorities regard how they want to accomplish their goals. Many experts, when asked what is required to accomplish their work will say "it depends" and launch into all the different situational considerations that feed into "how" they do "what" they do. Conceptually, goals can be characterized as "why", "what", or "how" type goals. In Figure 14 below, the "why" goal is the mission objective and answers the question "why is the human-agent team executing any work?" The "what" goals are the sequenced sub-goals that must be accomplished to achieve the "why" goal and answer the question "what must be done to execute the work?" Finally, in between the "why" and "what" are the "how" goals which express optimization considerations and answer the question "how should the work be done?" These goals, termed quality goals (Sterling & Taveter, 2009), are postulated to be correlated to the situation rather than specific elements of the task work.

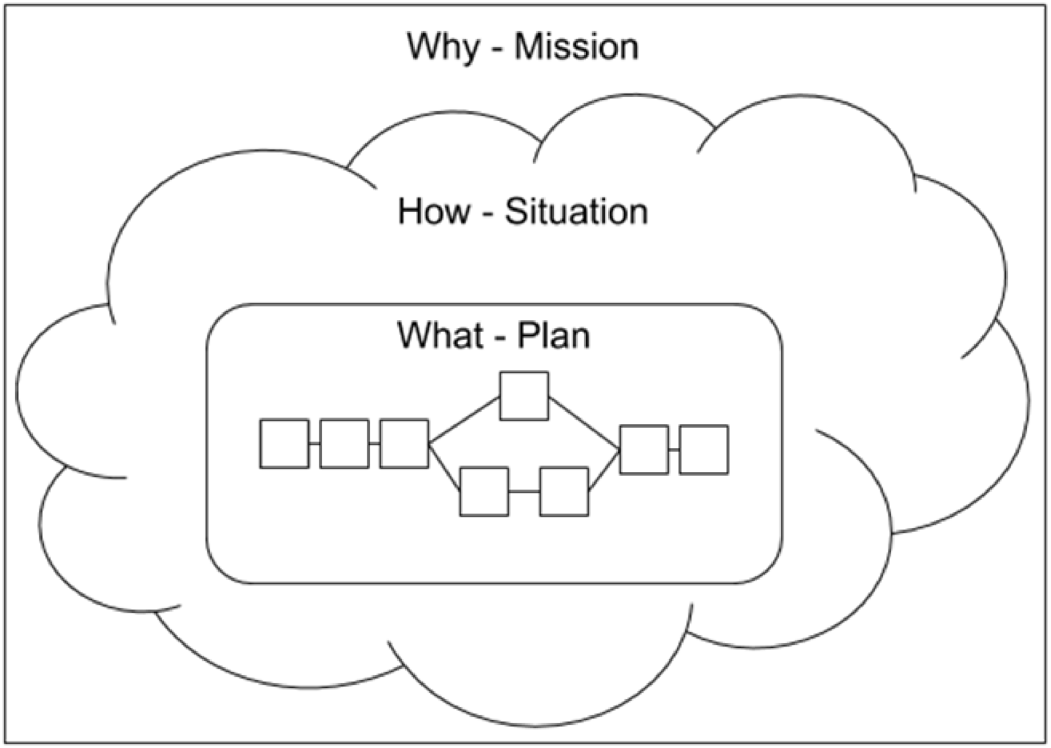


Figure 14. Relationship between “Why”-“What”-“How” based on Goal Decomposition

Since it focuses on quality goals associated with work, we start the design of the intent model with a cognitive analysis of the work to understand the operator’s perspective. A Goal Directed Task Analysis (GDTA) efficiently captures the environment, the work, and the systems considerations necessary to understand the critical priorities (Bolstad et al., 2002). We extended the GDTA to represent quality goals and to trace the information requirements of an operator to individual data elements which are observable to an AIA capable of estimating human intent.

A GDTA forms an acyclic directed graph in which functional goals are decomposed into their constituent sub-goals. These functional goals have decisions associated with them which entail information requirements to resolve. Thus, goals trace to decisions which trace to information requirements. The source data for a GDTA are

documentation on the domain (i.e. the environment, work, and systems) and interviews with subject matter experts (SMEs). This allows the entire domain, i.e., “why”, “what”, and “how”, to be captured. The meta-model for our GDTA is provided in Figure 15 as a Systems Modeling Language (SysML) Block Definition Diagram (BDD). The core GDTA elements are in white. The fuchsia blocks are the Operationalized Intent extensions and the light blue blocks are the data model elements. For further details on the intent model development see (Schneider, Miller, & McGuirl, 2020).

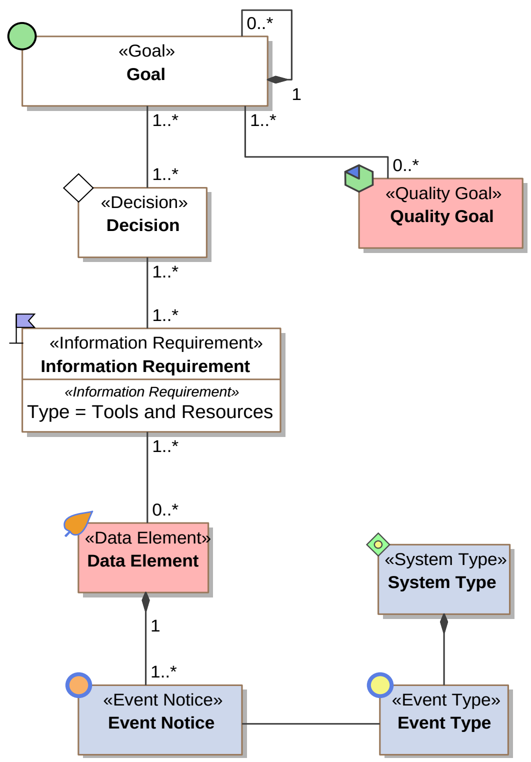


Figure 15. Extended GDTA Meta-Model

Data Element Identification

The GDTA information elements are the concepts, techniques and analytical results which enable decision making (Humphrey & Adams, 2011). However, data must be processed to become information (M. Chen et al., 2009). In complex data driven

systems, like aviation or nuclear power, the system itself is likely to have access to the same or more data than the operators. For intent estimation, the information requirements are traced to data elements, i.e. objects, which are present in the system and visible to the AIAs. This process provides a method to identify the relevant data in the system and understand how to transform it into information for the estimation algorithm. However, properly capturing situational data presents a significant challenge.

Event Notice Definition

Situational data is not only voluminous and high dimensional, the shape of the dimensions vary significantly. For our system, a remotely piloted aircraft (RPA) simulator, there were hundreds of possible data elements available. Aircraft state, operator interface information, weather, mission plans, are all well-defined with expected variability as there was only a single aircraft, single sensor, and single operator interface. The data model becomes more complex when there are an unknown, and possibly larger number of entities. In the current simulation, these entities include airspace restrictions, ground and air tracks, or chat messages in which data elements must be recorded in relationship to their parent entity. To capture this data the Core Situation Awareness (SAW) Ontology was used to capture situational data in the simulator (Matheus et al., 2003). Using the concept of an event notice, we defined timestamped data element logs associated with specific systems. Each event notice contains a description, type, system, up to six parameters, and two positions, see Figure 16 for example. The number of parameters and positions used, along with their definitions are uniquely defined for each event notice. Some notices used only one or two parameters, others used all six and both positions. The size of the event notice was determined by examining the data to be

logged. For example, the aircraft's positioning is required information for many of the decisions in the GDTA. Figure 16 illustrates the relationships between aircraft (A/C) positioning and the data elements that factor into an example information requirement as a SysML BDD. The *A/C Motion* data element is captured in the *a/c notice* event notice. This notice is an *aircraft_moved* type event which is associated with *RPA* type systems. The simulator was able to handle multiple RPA control, which the data architecture was required to support. In practice, since there was only one RPA in the simulation, only information for that system was logged. The data architecture included nine other event notice types associated with the RPA system and event notices associated with the other data elements. However, these data elements are not depicted here for clarity.

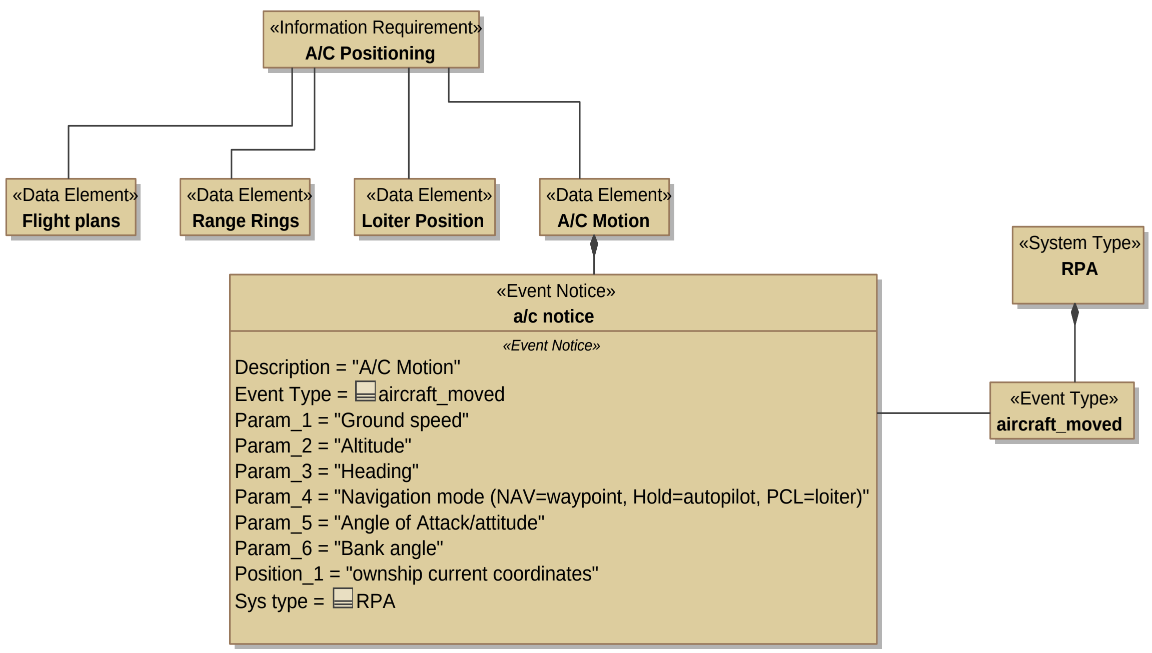


Figure 16. Example of Data Model

In the Core SAW ontology, an event notice is logged every time something causes "sensors to transmit new information" (Matheus et al., 2003). In our application, certain data elements were updated multiple times a second while others were initiated and never

updated. To accommodate this diversity without generating large volumes of unnecessary data, some event notices were logged at 2 sec intervals and others were logged asynchronously.

Study Summary

With a means to capture intent and the concurrent situation we investigated intent in simulated operations with experienced unmanned aircraft pilots. The details of the study can be found in other papers (Schneider, Miller, & McGuirl, 2020). The key elements of this study are summarized here to aid understanding. Twenty-one trials were performed by seven experienced operators, with each operator completing three trials. The three trials were designed to be dynamic and challenging with the operator's ranking the quality goals elicited at specific, situationally determined, points during the trial. The total dataset included over 115,000 event notices and 131 elicitations covering approximately 10.3 hours over 21 trials.

Estimation Context

To evaluate the differences between quality goal rankings we use a metric termed divergence (Schneider, Miller, & McGuirl, 2020) which is an exact, normalized version of the Spearman Footrule (Diaconis & Graham, 1977) given in (2).

$$Divergence = \frac{2\sum_{i=1}^{n}|R_{A,i}-R_{B,i}|}{n^2-(n \bmod 2)} \qquad (2)$$

where n is the number of quality goals in the hierarchy, $R_{A,i}$ is the rank of the i^{th} goal in the first set, and $R_{B,i}$ is the rank of the i^{th} goal in the second set. The divergence metric provides an unweighted, absolute distance between two sets of rankings which ranges from zero to one in discrete steps. We also defined a heuristic threshold which is

the minimum divergence at which the top ranked goal of one set is the bottom ranked goal of the comparison set with all other rankings being identical. For the nine goal model of this study, that threshold is 0.4. It is possible to achieve a higher divergence without reversing the top and bottom rankings. In comparing the operators divergence we demonstrated that, once a group of three high divergence trials were removed from the data set, the distribution of divergence across operators in the same situations were not significantly different from the top/bottom threshold (Schneider, Miller, & McGuirl, 2020). The high divergence trials are included in the dataset applied in this research.

Translation and Labelling

While we present a specific estimation algorithm in this paper, the data collection process is algorithm agnostic. An implemented real-time estimator would have to contend with multiple subsystems feeding in event data, potentially at multiple reliability levels. All these data must be translated from their source form into a more informational context. We use the example of the Federated Relational Database System as a means to process the data into cohesively understood information (Blanco, Illarramendi, & Goni, 1994). Starting with the raw event notices, the data are translated into a common meaning and then integrated into a situation vector for the specific algorithm we implemented.

Situation Vector Development

Intent estimation is formulated as a label ranking problem. A label ranking algorithm receives a vector describing the instance to be estimated and produces a ranking of defined labels that best fits the instance. For the present research, the situation vector contained a total of 113 features. These covered operator interface interaction,

aircraft state, sensor information, navigation, weather, and key spatial references. The event notices contained the simulation names for various entities, e.g. chat room call signs, reference point code names, etc. Each was translated into a generic form focused on the role that name represented, e.g. call signs were translated to the role that entity played like “customer” or “reference point 1.” All spatial coordinates were endogenously referenced to the aircraft position or sensor point of interest (SPI) as a range and bearing. These two points can be equally important to the RPA operator’s decision making. In the resulting vector, all entities, including entities like reference points, had their representation change every time the aircraft or SPI changed.

To represent the variety of transient entities in the simulation, i.e. airspace restrictions, air tracks, and ground tracks, we employed a state space representation which related the entities based on proximity to the aircraft and sensor focus point. The method developed by Bindewald et al. defines position and relevance based on an ownship location and a destination (Bindewald, Peterson, & Miller, 2017). For an RPA the two foci of the state space are the aircraft position and SPI location. We define four zones around each such that zone one is the quadrant that faces from the aircraft to the SPI, zone four faces from the SPI to the aircraft with zones two and three being to each side. A score was developed for each type of entity in each zone representing the quantity and proximity for a total of 24 features.

There were four additional special features, which were temporally correlated with the simulation, explicit communication, and operator identification. The *Fuel Level* and *Time to Bingo* features represented the percentage of fuel remaining and the time until a certain fuel threshold was reached. Each trial began with full fuel tanks and the

fuel burn rate never significantly changed during the simulation, so these two features provide a sense of sequencing to the situation vector. They are not correlated to the events of the trial or quality goal rankings, but they provide a sense of time. Operationalized Intent seeks to minimize explicit communication about ranking of quality goals. So to represent minimum explicit communication, it is possible to envision a system where the highest priority quality goal name is provided as a feature to the algorithm. This is equivalent to the operator designating only their top ranked quality goal in an interface. Finally, the operator's two digit identifier was included as a feature to assess whether individual personalization could be effective at predicting intent. Various models were constructed in the current research which included or excluded selected special features to understand the utility of each of these features to the intent estimate.

To integrate event notices and form situation vectors temporal windows were used to compile event notices into updates to the situation vector. The initial situation vector is assembled using approximately the first 20 seconds of event notices, less than 1% of the total data available. This ensured that each trial had an independent starting point. The fastest periodic event notices updated every two seconds, so windows of 1, 2, 3, and 4 seconds were used to assemble the event notices. The trials were divided into window length segments and the last event notice of each type for each system was used to update the situation vector. This method obscures some event notice data for events that occur multiple times within the window. However, the average percent difference between one situation vector and the next varies between 21.0% and 22.4% for window sizes of 2-4 seconds. Window size of 1 second has a mean of 16% difference, but this is

largely due to the fact that the aircraft position only updates every other situation vector. Given that there are only ~37,000 seconds available in the dataset, increasing the window size to 4 seconds reduces the volume of data to less than 7,000 situation vectors. Using this process, the raw event notices are translated to a common schema and then integrated into the situation vector representation for use with the machine learning algorithm. This requires labeling of each vector with goal ranks to provide a complete data set for supervised machine learning.

Labeling Assumptions

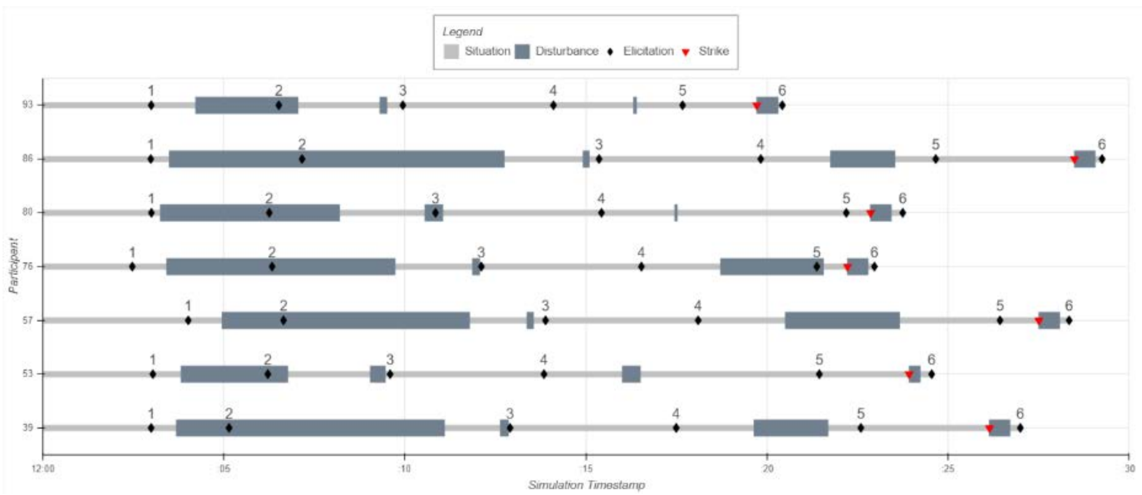


Figure 17. Timeline of Trial 1

As described above, “how” type intent is postulated to be situationally stable, such that until a situation is disturbed, the intent remains approximately unchanged. To label the situation vectors the situations and disturbances for each operator and each trial were identified based on their individual data. The intent elicitation for the situation most closely resembling that condition was then applied as the ranking for the labels. For example, in Trial 1, depicted in Figure 17, the first elicitation (black diamond) occurs during the first situation, surveilling a compound. The first disturbance (dark gray

section) was an unplanned new airspace restriction which required some navigational re-planning. Once that was completed, the operator resumed their task of surveilling the compound. This enabled the labeling of the first and second situations with the same elicitation ranking. Some elicitations, like elicitation 3, did not occur during the disturbance they capture, but immediately after. Using a situational labeling scheme allowed for proper labeling of those situations.

Splitting

The fully labeled dataset was split 75/25 between testing and training. To ensure independence whole trials were split with 16 trials for training and 5 for testing. The random draw of test trials had to meet two criteria, it had to include one of each trial type (i.e. Trial 1, 2, 3) and at least three different operators. This forced the test set to be nominally representative of the training set.

Estimation Algorithm

Operationalized Intent Estimator – Random Forest (OIE-RF)

The OIE-RF algorithm is based on the Label Ranking with a Random Forest (LRRF) design by Zhou and Qiu (Zhou & Qiu, 2018). For details of LRRF see the comprehensive summary of the algorithm and its performance in their manuscript. In this section we provide a brief overview of their algorithm and highlight deviations from the base design.

LRRF is a random forest algorithm, training a number of decision trees (the forest size) each on a bootstrapped, random subset of the training dataset. The predictions of these trees are then aggregated to produce the overall model prediction. Decision trees in

this context are binary trees used for classification tasks. Each node of a decision tree splits the domain of the input data into two subsets based on a particular feature in the situation vector. Each subset is then further split by the children of the node, until a leaf node is reached. Using this smallest subset represented by a leaf node, we can use the subset class labels to find a predicted class. In our application, this predicted class includes the label ranking, i.e., a predicted ranking.

The tree construction algorithm is the same for each tree in the forest. Each tree is first assigned a random subset of the dataset. At each step, a subset of features is selected, and split the training data based on the feature that would produce maximum information gain. This process produces a measure of change in entropy as a result of a split. To calculate entropy, LRRF employs the Top Label as a Class method, treating the highest ranked label of each ranking as the class of that sample.

Our implementation of OIE-RF uses three methods of finding the optimal split, based on the data type of the feature. The first, *Threshold Split* from the original LRRF algorithm operates on quantitative and Boolean data by finding an optimal threshold such that for each instance x_i of feature x:

$$\begin{cases} x_i \geq threshold & : \ x_i \text{ in left subset} \\ x_i < threshold & : \ x_i \text{ in right subset} \end{cases}$$

Categorical Split operates on text data and considers each unique class in a categorical feature, and partitions the dataset based on equality to that class. Specifically, it finds the optimal class such that for each x_i:

$$\begin{cases} x_i = class & : \ x_i \text{ in left subset} \\ x_i \neq class & : \ x_i \text{ in right subset} \end{cases}$$

Certain features in the collected dataset contained null values. These are not missing data, but rather represent features which are not pertinent in that situation. As an example, the aircraft has a “loiter” navigation mode. If the operator is flying the aircraft directly, the values of the features describing the loiter mode are designated null, which is correct because there is no loiter mode data to be described. For categorical data this presented no issue, as "Null" could be treated as a unique class. For quantitative data, the addition of a *Null Split* was necessary as null are not valid for greater-than or less-than comparisons. To remedy this, we first perform a Null Split on the quantitative features if null values are present. This simply places all null values into the left subset, and non-null values into the right. For each x_i:

$$\begin{cases} x_i = Null & : \ x_i \text{ in left subset} \\ x_i \neq Null & : \ x_i \text{ in right subset} \end{cases}$$

Once the optimal split for each feature is found, the feature resulting in the most information gain is designated the split feature for the node. The algorithm is then repeated on the left and right subsets, recursively forming the tree on smaller subsets of the training dataset. The halting rule designates leaf nodes if one of the following conditions are met: 1) the maximum specified depth is reached, 2) the number of unique classes, i.e., top labels, is less than 2, or 3) the number of training samples in the left or right subset is 0.

On a leaf node, we store the subset training rankings represented by that leaf. To aggregate the rankings, LRRF uses a two stage, generalized Borda count. Borda count is a consensus voting method which provides a means of producing an unobserved ranking. This is imperative since there are 9! permutations (over 362,000) of the goal hierarchy

and the dataset contains 131. The rankings are aggregated at the leaf node during training and across the trees during estimation testing. OIE-RF is implemented in Matlab 2020a.

Hyperparameter Tuning

OIE-RF has three hyperparameters of forest size, tree depth, and window size that require tuning. For all random forest algorithms, the forest size and the tree depth are critical hyperparameters. The window size for situation vector aggregation discussed earlier is the third hyperparameter as it influences data set size and granularity. We investigated four forest sizes, i.e., 25, 50, 75 and 100 trees, three tree depths, i.e., 5, 7, and 9, and four window sizes, i.e., 1, 2, 3, and 4 s. The factorial combination of which resulted in 48 permutations. These conditions were reduced to 35 permutations by comparing the maximum number of leaf nodes for a forest with the number of situation vectors in the training set. This eliminated conditions where the algorithm was likely to overfit (e.g. deep depth, large forest, on a small dataset based on window size), or underfit (e.g. shallow depth, small forest with a very large data set based on window size).

Using a K-fold cross validation technique to examine the 35 permutations, it became apparent that forest size and window size were positively correlated to computation time while tree depth was negatively correlated to divergence. However, the K-fold cross validation was determined to provide an inappropriate criteria for tuning of the parameters as neighboring labeled situation vectors were not independent of one another, which resulted in unrealistic divergence performance. To recheck the hyperparameter performance we ran four conditions, given in Table 10, which captured all the levels of the hyper parameters using 14 training trials and 2 independent cross

validation trials. To assess the quality of the hyperparameter models we compared mean divergence performance, mean leaf node entropy, the coverage ratio of the number of features used by the model to the total number, and the mean of the tree fullness or percentage of leaf nodes at the maximum depth. Desirable conditions resulted in low divergence, low mean entropy, and high coverage. However, a balance must be struck between high coverage with fullness. While lower fullness indicates that a tree is reaching the halting condition earlier, a low fullness indicates that the dataset is being memorized by the tree. Forest size had a minimal impact on feature coverage and window size did not have a strong impact. Divergence had a strong negative correlation with tree depth, but the decrease between a depth of 7 and 9 was minimal. The fullness of a depth 7 model was approximately 70%, but dropped to 25% at a depth of 9, indicating that depth of 9 was overfitting to the data. The conclusion was that a window size of 4 seconds at a tree depth of 9 layers in a forest of 25 trees provided a model that was unlikely to overfit the data, but deliver accurate estimates rapidly.

Table 10. Hyperparameter Study Permutations (selected parameters shaded)

Window (sec)	**Tree Depth (layers)**	**Forest Size (trees)**
4	7	25
1	7	50
3	9	75
2	5	100

Estimation Models

To assess the impact of the four special features listed in Table 11, five models were trained. Models 1 and 2 contain only situational data and examine how accurate strictly implicit estimation can be. Explicit communication of the operator’s highest

priority or top ranked quality goal is tested by Models 3 and 5. Model 1 and 5 do not include the temporally correlated variables to assess their impact on the model. Model 4 contains all features including the operator's identity which allows for individually tailored estimates. In the analysis, *Highest Priority* and *Operator ID* are included as features, but do not entail any restructuring of the algorithm to force the highest ranked quality goal to be listed as rank 1 or to only consider data from a single operator.

Table 11. Model Feature Inclusion

Model	Fuel Level	Bingo	Highest Priority	Operator Id
Model 1 SS				
Model 2 SD	X	X		
Model 3 EC	X	X	X	
Model 4 IT	X	X	X	X
Model 5 ES			X	

Results

Accuracy testing is divided into two sets of tests. The static test observes the divergence between the estimator and the situation vectors closest to the elicitation time in the test set. These are minimally impacted by the postulate of intent and situational stability. This is a pure accuracy test. The dynamic test includes all situation vectors in the test set and observes the sequential stability and overall accuracy of the model over time. From this we can observe the estimator's sensitivity to situation change and if the estimator converges over time to a lower divergence estimate. Finally, using the data from these two tests we can compare the estimation models and rank their overall performance.

Static Test Results

The static test evaluated the divergence between the estimate and the actual labels for the two situation vectors closest to the elicitations. The test set of 5 trials provided 56 total test points. The median divergence across the five models was between 0.4 and 0.45. A Kruskal-Wallis test of the distributions found no significant difference between the estimates and the top/bottom threshold for Models 1-4, with Model 5 providing divergence that was significantly higher than the threshold (α=0.05). Figure 18 illustrates the resulting divergence for each model in a jitter plot with all data points in the same vertical band having the same value. The medians and interquartile range (IQR) are included as a distributional reference.

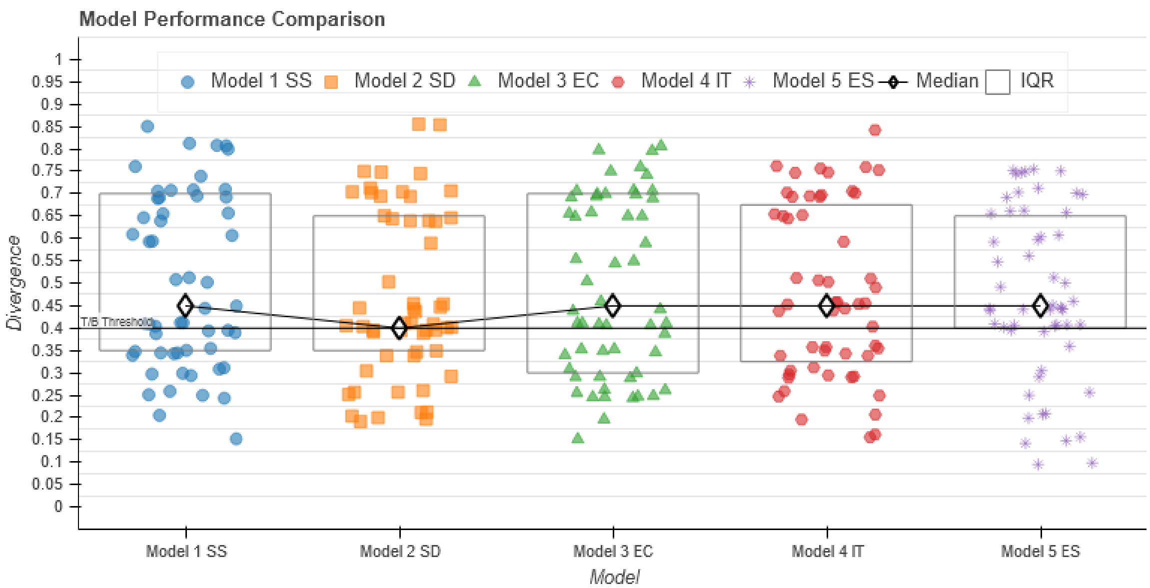


Figure 18. Static Test Results by Model

There is an apparent bimodality to the divergence of some models with a high divergence grouping around 0.7 and a low divergence grouping around 0.4. If the same data are plotted against the five test trials the cause of this bimodal distribution becomes

clear. The trials listed on the horizontal axis of Figure 19 correspond to an operator id (e.g. P57) and a trial number (e.g. T2). The color indicates the model which estimated the intent. Trial P57_T2 has the highest divergence with P53_T3 having the second highest median divergence. This is not exclusively the result of poor performance on the part of the estimator. In the initial study data analysis, the highest divergence operator for each trial was identified in the analysis. The highest divergence operator for each trial consistently ranked the goals differently from the rest of the operators and so has high divergence with everyone. Of those three trials from the 21 in the dataset, two were randomly included in the test set, P57_T2 and P53_T3. These operators held a different perspective, not wrong only different, from the other six and so it is unsurprising that the estimator has a similar high divergence within these two trials.

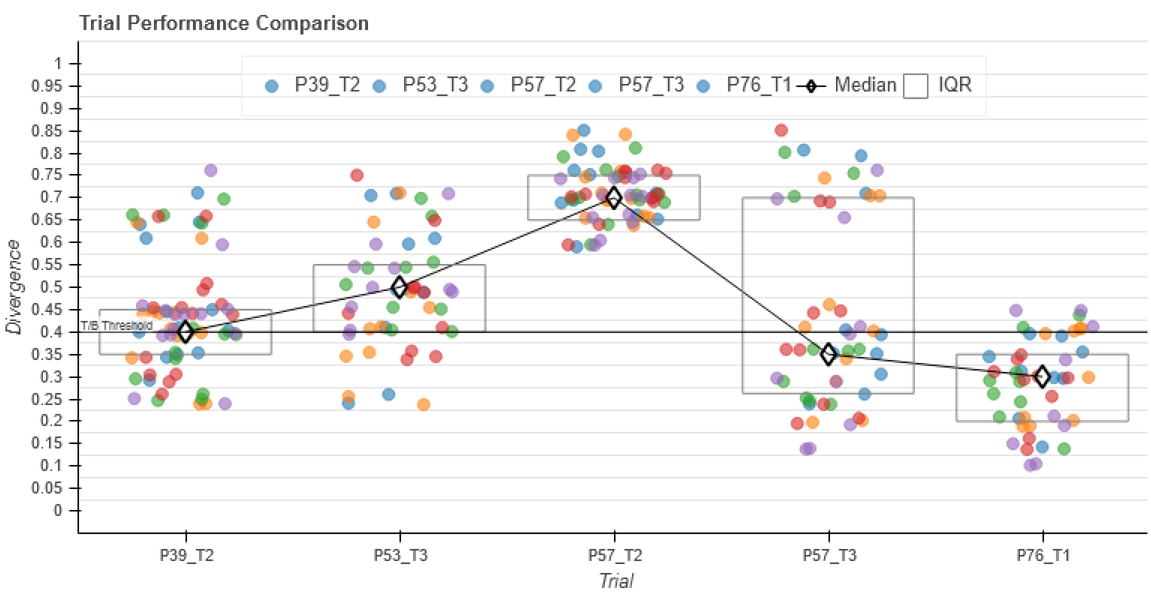


Figure 19. Static Test Results by Test Trial

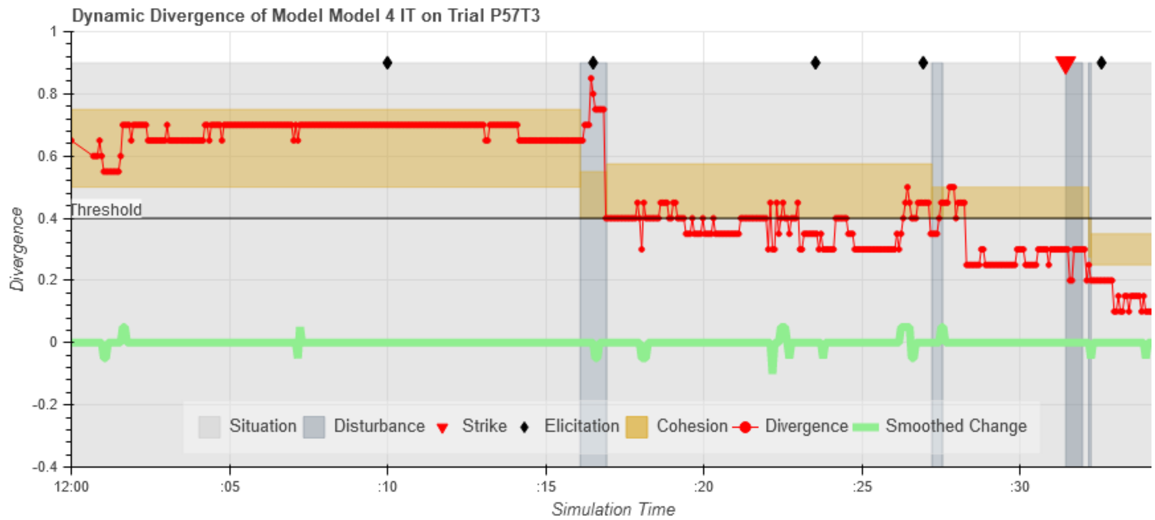
Dynamic Test Results

The dynamic test used all the situation vectors of the test trials in sequence (n=2,353). This enables time series analysis of the data to understand the sequential stability of the estimator and the sensitivity of the estimator to situational change. We define stability as the divergence shifting slowly without significant variation within a situation. A three-period moving median is used to assess stability. By comparing the divergence change over time with the situation changes in the trial we can observe the sensitivity of the algorithm to situational change and if there is any evidence of convergence. There are twenty-five charts like the two in Figure 20 to capture results from each of the 5 models and 5 test trials. The light gray portions are the situations, the dark gray are the disturbances with the black diamonds at the top marking the elicitations. The tan cohesion bands represent the interquartile range of the distribution of divergence among the operators for that situation. It provides a comparison of the estimator divergence with the total population cohesion distribution. The green line is the smoothed change in divergence with the red line being the divergence between the estimate and the elicited intent for that situation vector.

The top panel in Figure 7 represents the data from trial 3 as compared to the results from Model 4 plot. As shown in this panel, the estimate of divergence monotonically decreases with time and for the latter half of the trial is largely below the cohesion band. The plot illustrates that the estimator did not observe the first disturbance which is why the divergence spikes during this disturbance. This disturbance is a lost link event where the operator loses command of the aircraft. A review of the data set revealed that link status information was not captured in the situation vector which may explain

the insensitivity of the estimator to this disturbance. However, after the initial spike the divergence decreases which demonstrates that the algorithm is converging to a more similar ranking.

The lower plot in Figure 7 is trial 2 for Model 4. Insensitivity is again observed at the first disturbance where there is a jump in divergence with no attempt to correct. This disturbance was a generator failure event that the situation vector fully captured. However, this disturbance demonstrates that in certain conditions the estimator was not successful in detecting a change in intent, producing a significant increase in divergence over the time period relevant to the change in resulting operator intent. The fourth situation (fourth light gray region in the bottom panel of Figure 7), demonstrates convergence where the estimator refines the quality goal ranking as time progresses and over five minutes reduces the divergence by 50% from 0.45 to 0.2 during the situation.



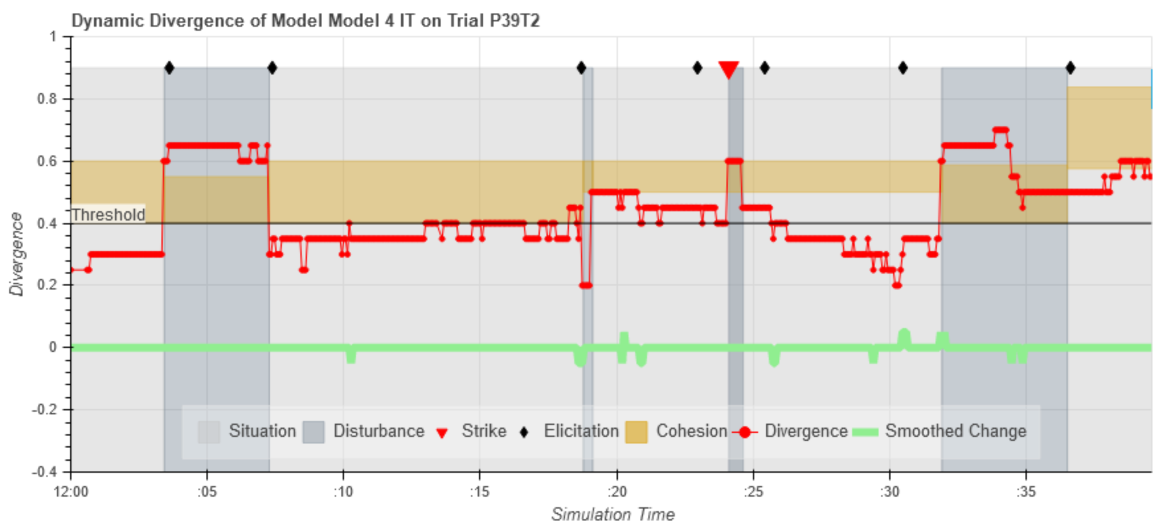


Figure 20. Example Dynamic Test Results for Trials 2 and 3

There are some spikes and areas of instability in the smoothed change line which represents times when the estimator is changing the rankings rapidly. Observing the distribution of the raw and smoothed change in Figure 21, it is clear that the algorithm remains stable with zero change over 80% of the estimates. The accuracy of the dynamic test is slightly higher than the static test with a median of 0.5. However, the major driver

of the poor accuracy is trial P57_T2, the exclusion of which reduces the median to 0.4, commensurate with the static test.

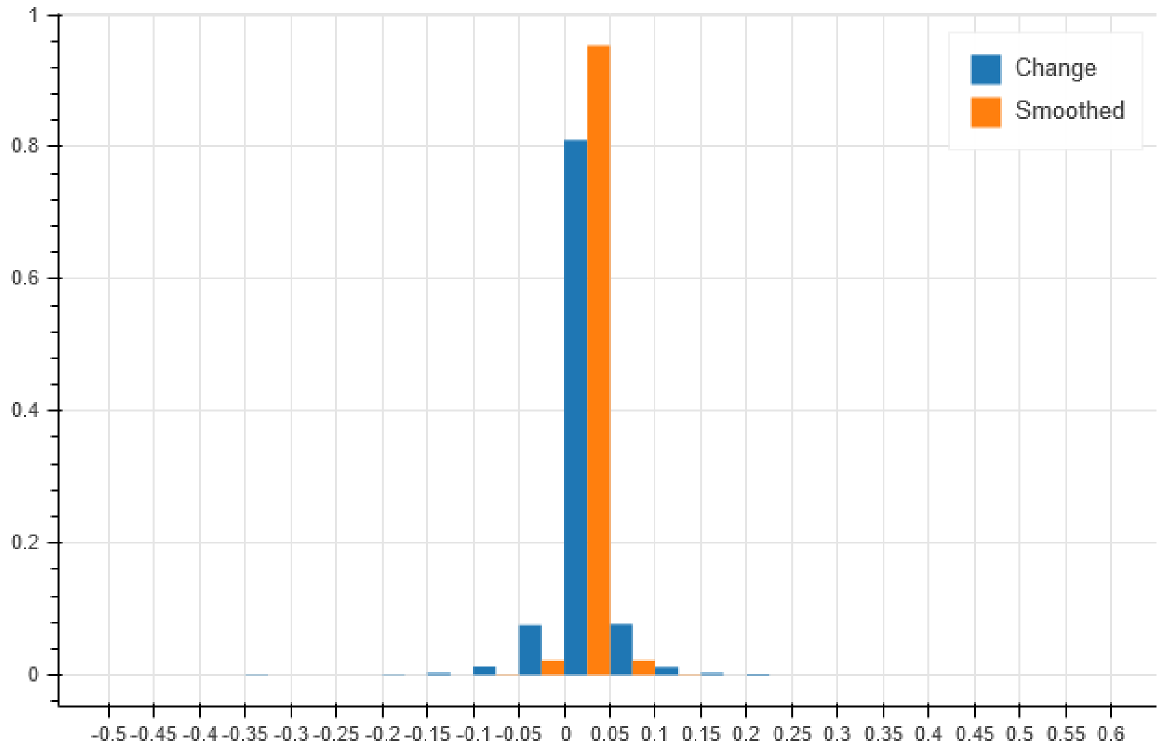


Figure 21. Histogram of Sequential Divergence Change Demonstrating Estimator Stability

Model Comparison

A Kruskal-Wallis test between the distributions of the static accuracy data by model were not significantly different (p=0.97). This indicates that there is not much difference between the accuracy of one model vs another. However, the dynamic test results provided an alternate means of comparing model performance. Additionally, by examining the random forest we can leverage the hyperparameter metrics and understand the structure of the model to assess the influential features.

Performance

We define a situation baseline as the interquartile range of the divergence distribution from the operator responses within a situation. The divergence distribution is calculated through an all vs all distance comparison of responses. This is represented by the tan-colored band shown in Figure 20. To judge model performance, we compare the percent of estimates which have a divergence less than, equal to, or greater than the situation baseline.

For an estimator with the same accuracy as the situation baseline there should be approximately 25% above, 50% within, and 25% below the situation baseline. Instead the mean percentages of the models ranged from 16.4%-21.9% above, 42.4%-54.8% within, and 28.9%-40.8% below. The performance of Model 4 is provided in Figure 22 as an example with green illustrating the percentage of estimates below the situation baseline, yellow within the baseline and red above the baseline. This again demonstrates the challenges of estimating high divergence trials like P53_T3 and P57_T2 which have the least green and much of the red and yellow regions.

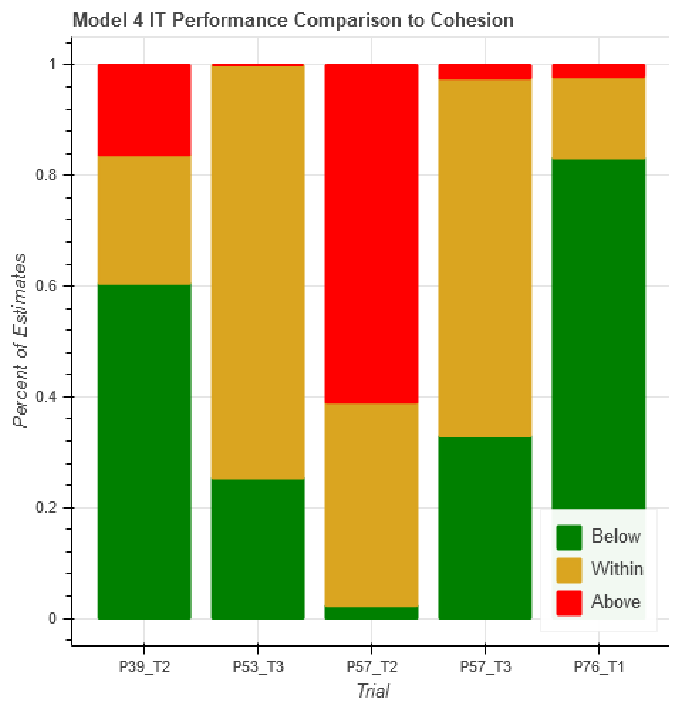


Figure 22. Model 4 Performance Compared to the Situation Baseline

However, it is better for the algorithm to provide estimates with divergence within the situation baseline and better to be below this baseline than within. We established two ratio metrics for model comparison. The first is the percentage of estimates above the situational baseline versus within and below. The second is the percentage of estimates within the situation baseline versus below the baseline. Using these two metrics we then calculated a ranked sum score for each model. In this case a low score is best, Table 12 provides the order of performance with Model 4 being the best. Clearly having more personalized information with explicit communication provides a better estimate of quality goal rankings. Interestingly, the temporally correlated features also seem to provide a performance boost since the top three models include them and the bottom two do not. Model 4 had an average of 16.4% above, 42.8% within, and 40.8% below the situation baseline. On average there were nearly three times

as many estimates within and below than above and over four times as many within as below. Comparing the accuracy of the estimator to the situation baseline strongly demonstrates that this algorithm can estimate intent more accurately than the situation baseline.

Table 12. Performance Ranking of Models

Model	Score
Model 4 IT	0.418
Model 3 EC	0.526
Model 2 SD	0.530
Model 5 ES	0.552
Model 1 SS	0.574

Structure

In addition to the performance comparison between the models, random forest provides an observable model which allows examination of the structure to understand the critical features. Model 4 has the highest feature coverage at 97%, lowest average entropy at 0.78, and lowest fullness at 66%. This indicates that Model 4, while using all but four features, is efficient at driving to high purity leaf nodes rapidly. The explicit communication and operator information are key to this structural efficiency. The split feature for each node in a tree is logged, Figure 23 provides a Pareto chart of features used in the model. The color code designates features that are close to each other in the feature vector and likely to be related. The five most frequently split on features are the *Highest Priority*, *Participant* (i.e. Operator ID), *Closest Restriction Range from Ownship*, *Fuel Level*, and *Time to Bingo*. This corroborates the performance conclusions regarding the utility of the four special features.

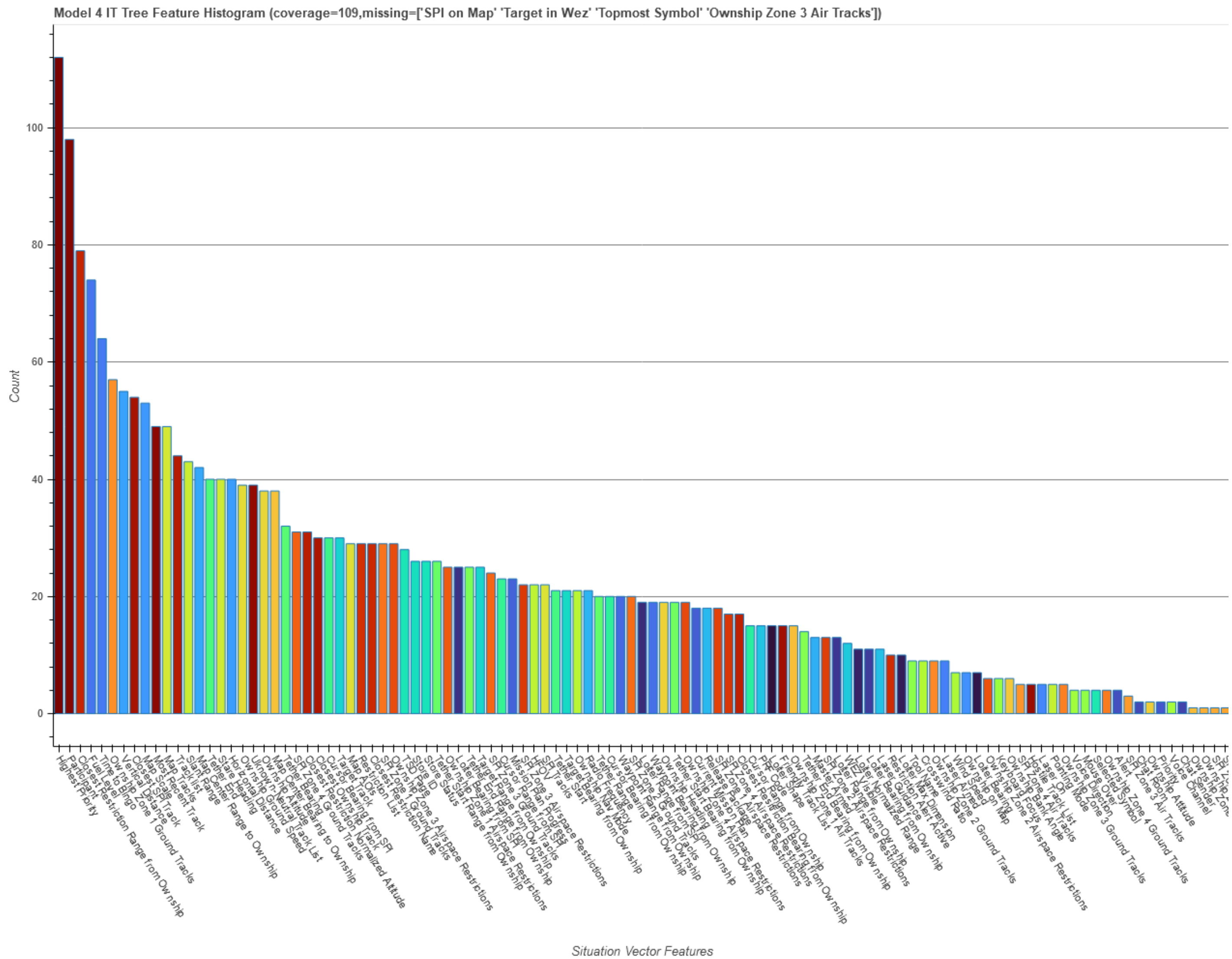


Figure 23. Pareto Chart of Feature Split Count in Model 4

Discussion

The desired objective of Operationalized Intent is a role-centric, rather than operator specific, estimation technique. The dynamic performance data corroborate the previous findings that knowledge of, and communication with, the operator being estimated improves estimation accuracy. The communication related features (e.g. voice channel, chat sender, etc.) are not frequently used which indicates that sparse data are not represented effectively in the OIE-RF algorithm. Alternative algorithms and data representation methods designed to handle sparse and continuous data together may prove more effective.

The dynamic test demonstrates that stable estimates of intent can be provided, but must be balanced with situational sensitivity. A key benefit of graph based associate systems discussed in the literature is the ability to rapidly shift situational understanding, which provides improved sensitivity (Geddes & Lizza, 2001). The stability of OIE-RF may allow for improved coordination with the operator since providing the operator's top ranked quality goal is retained by a stable algorithm. In addition, a more stable algorithm inherently displays more predictable behavior to the operator which encourages teaming (M. Johnson et al., 2018).

It is important to note that the graph-based approach applied in associate systems requires the expert construction of the knowledge graph, which provides an implicit representation of both "what" and "how" intent. As discussed in the literature, this implicit representation reduces the operator's ability to observe, predict, and direct the behavior of automation within the system (Miller, 2017). The current approach provides

an explicit representation of the system's understanding of goal priorities relating to "how" type intent which can be shared and altered by the operator. As illustrated, including the operator's input to the algorithm in the form of the goal with the highest ranking reduces the divergence of the algorithm with the intent of the operator, permitting changes based upon operator input.

In the truest test of accuracy, OIE-RF is meeting the heuristic threshold. With divergence ranging from 0.15 to 0.85 it resembles the overall cohesion of the operators which ranged from 0.1 to 0.9. While algorithmic and estimation techniques can be improved, increased operator training and refinement of the intent model is also likely to contribute to accuracy improvements. At the opening of Trial 3, illustrated in the top panel of Figure 20, the operators were faced with an initially ambiguous situation that may not be frequently encountered (i.e. on the wrong side of a storm with a hard deadline to support a convoy), and so the variability in the operators' intent models during this situation is large. However, towards the end of Trial 3 the operators encountered an all too familiar situation, supporting friendly troops in contact with an adversary then pursing fleeing adversaries. The divergence among operators is much lower (below threshold) in this case and the estimate has improved accuracy. The more operators leverage common mental models, the more useful an intent estimator can become.

For machine learning algorithms, much of the intelligence is imbued in the data representation. Leveraging the same cognitive engineering analysis which produced the quality goals, we were able to identify the specific data elements to record from the simulator and their relevance to the operator's decisions. Using the Core SAW Ontology (Matheus et al., 2003) and the Federated Relational Database framework (Blanco et al.,

1994), the raw data were captured in an efficient manner, translated into a common syntax and reference frame, and then integrated into a situation vector that provided the estimator an informational representation. This data representation process is designed to be scalable across more complex systems and can be ingested from modular subsystems. The total process instills rigor and justifies which data are used, and how they are represented to avoid algorithms identifying inappropriate patterns on which to base their estimates.

OIE-RF trained the five models each in about 20 minutes on a quad-core Intel Xenon processor and could predict an entire trial's worth of data in a few seconds, significantly faster than the window size. This computation efficiency demonstrates that an operational system is likely to be able to operate in real time within a human-agent team.

Conclusion

OIE-RF is a single step machine learning algorithm which could be improved by additional steps and accompanying algorithms. Much of this research was undertaken in advance of the release of TensorFlow-Ranking library (Pasumarthi et al., 2019) which is a deep learning library for the label ranking problem. Investigating these and other ensemble methods for intent estimation are likely to improve accuracy, sensitivity, and stability.

As an initial investigation into the feasibility of modeling and estimating intent, this research clearly indicates that quality goal rankings associated with "how" based intent can be explicitly modeled and estimated accurately from situational data and

minimized explicit communication. Explicit communication and knowledge of the operator proved useful to improving accuracy. Estimated intent can be shared with a team of AIAs to provide them a means of understanding their human operator teammates and coordinate their actions more effectively. This research demonstrates that intent estimation is feasible and opens a new avenue to improving coordination and performance of human-agent teams.

VI. Conclusion

Research Conclusions

While each of the preceding chapters provided conclusions, here we summarize the entire research as it addresses the research questions.

Through a synthesis of coordination, intent, and mental models, we identified the relevance of “how” type intent to synchronize action across a human-agent team (HAT). Constrained by the operations tempo of high performing teams, leveraging their training and domain specific vocabulary, the Operationalized Intent theory provides a framework for designing a shared mental model. By structuring the representation as an ordered set of quality goals and a status of execution constraints, the intent model is computationally ingestible by AIAs. The quality goal hierarchy estimation is a label ranking problem for which ample computer science research efforts and algorithmic techniques exist. Situation based estimation of intent enables implicit coordination. With an explicit model of intent, the HAT members can directly coordinate. This coordination extends to the human operators which allows for explicit coordination of intent as well. Thus, Operationalized Intent provides a scalable framework for designing shared mental models regarding how tasks should be completed as a complete coordination mechanism.

To apply Operationalized Intent requires understanding the domain, the operators, and the AIAs. Beginning with a Goal Directed Task Analysis (GDTA) to capture the domain, the candidate quality goals, and data elements are identified. These are heuristically evaluated for suitability based on the AIAs. We demonstrated the method of leveraging the GDTA to develop a set of study trials to examine the effectiveness of the quality goals using dynamic and static evaluation methods. Identifying the situational

data key to estimating the quality goal ranking is derived from the GDTA as well. Our study establishes a scalable means of capturing, translating, and integrating that data into a form to support estimation of the quality goal ranking.

The study results indicate that the intent model was able to capture intent shift based on the situational disturbances. Overall use of the quality goals indicated that the rigorous development of the quality goal hierarchy and the trial design was effective. Using the divergence metric and situational comparison across participants, the study results indicate that the operators demonstrate a heuristically reasonable level of cohesion. Identifying high divergence contributors motivates a level of personalization in intent estimation to account for the variety in perspectives of different operators. It also indicates that further operator training may be useful to embed the intent mental model more effectively.

Leveraging an available label ranking algorithm, we demonstrated that, static estimates of intent can be provided with accuracy on par with the overall intent cohesion of the operator population. When evaluated dynamically, including explicit communication and operator identity representations, the estimator was able to outperform the situation baseline. It demonstrated temporal stability in the estimate and in many cases convergence over time.

Contributions

The primary contribution of this research is the Operationalized Intent framework. Due to the academic format of this dissertation, the theory and domain application process of Operationalized Intent is spread over three papers. The following section

provides a cohesive summary of Operationalized Intent with the subsequent section summarizing all the contributions of this research.

Operationalized Intent Theory and Domain Application Process Summary

Operationalized Intent theory is described in detail with exact definitions at the end of Chapter III. The purpose of Operationalized Intent is to enable implicit coordination between operators and functional AIAs by continuously estimating a model of the operator's "how" type intent. Figure 24 is a reproduction of Figure 4 from Chapter III which graphically depicts the ontology of Operationalized Intent.

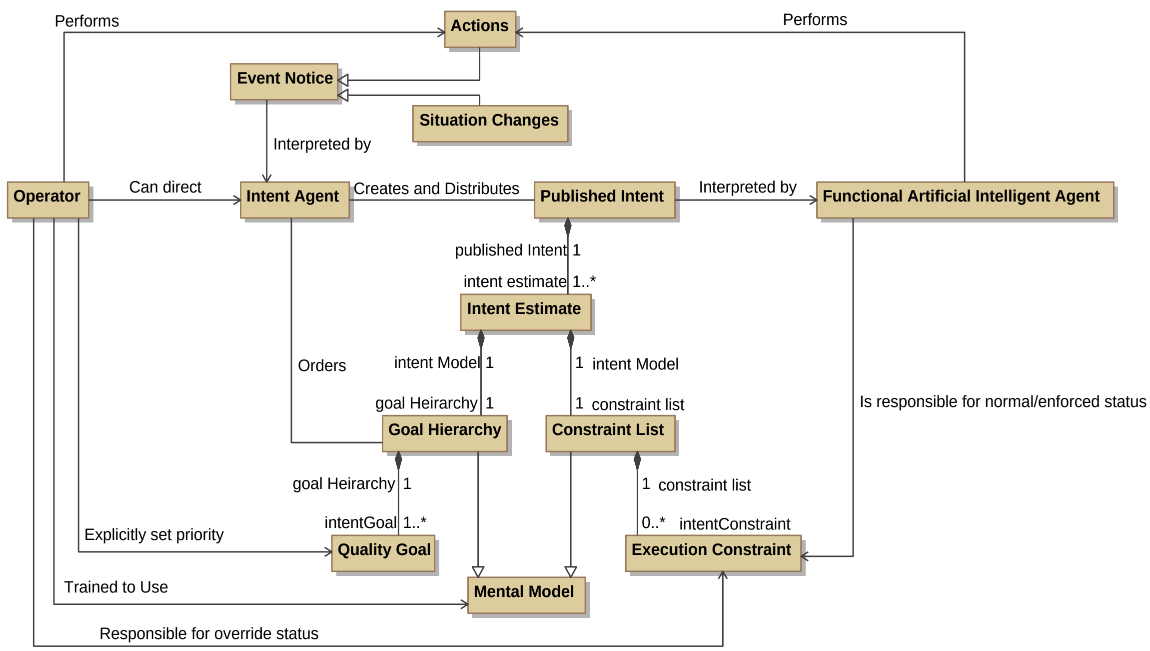


Figure 24. Ontology of Operationalized Intent

The intent model is composed of quality goals and execution constraints. Quality goals provide execution guidance to help functional AIAs shape the topology of their behavior trade space to find the most useful behavior to synchronize with the operator. While not sufficient to fully shape the behavior trade space, quality goals are one of many

inputs to an AIA's method selection process. The quality goals are defined during design and ranked in an ordinal hierarchy during execution. Execution constraints identify limits to AIA behavior trade spaces relevant to the understanding of the system by the operator and other AIAs. Hard constraints cannot be overridden, while soft constraints can be overridden by the operator. Execution constraints are defined at design time and change status during execution.

During operations, the quality goal hierarchy and execution constraint status list form the intent estimate. It is possible, although not necessary, to sequentially composite intent estimates of likely future intent to form a published intent. These future intent estimates would be based on mission plans and updated as the situation changes during execution. The published intent provides the functional AIAs an immediate and forecasted model of the operator's intent mental model.

The published intent is managed and updated by an AIA specifically designed to estimate intent. This agent is referred as the Intent Agent. By observing the actions of operators, functional agents, and situational changes in the system and environment, the intent agent estimates intent by ordering the quality goals into a hierarchy and compiling the status of the execution constraints set by the designated functional AIAs and the operator. An operator interface should be provided to directly communicate the operator's ranking of the quality goals into a hierarchy to facilitate explicit coordination.

For Operationalized Intent to serve as an effective shared mental model, it must be embedded in the cognition of the functional AIAs and the operators. In the case of functional AIAs, the designers include the intent model into the method selection process of their agents and test that it is producing functionally relevant differences in behavior.

Operators must be trained on the use of Operationalized Intent and practice using the model to embed it in their cognitive processes.

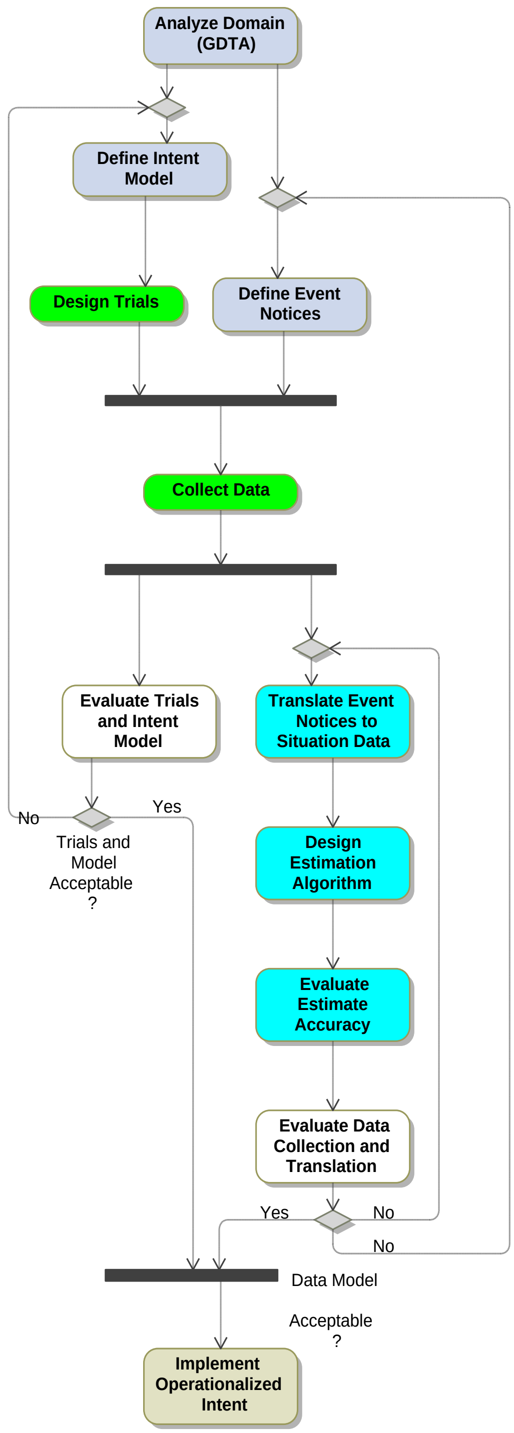


Figure 25: Domain Application Process Activity Diagram

To apply Operationalized Intent to a specific domain and system requires a thoughtful design process. This domain application process is addressed partially in Chapter IV, intent model construction and trial design, and Chapter V, event notice design, translation, and estimation. Figure 25 provides a SysML activity diagram which describes the domain application process. The domain application process occurs in four phases: analyze (gray), study (green), estimate (blue), and evaluate (white), which are color coded in the activity diagram. Many of the elements discussed below are captured in Figure 26 as a meta-model for intent analysis.

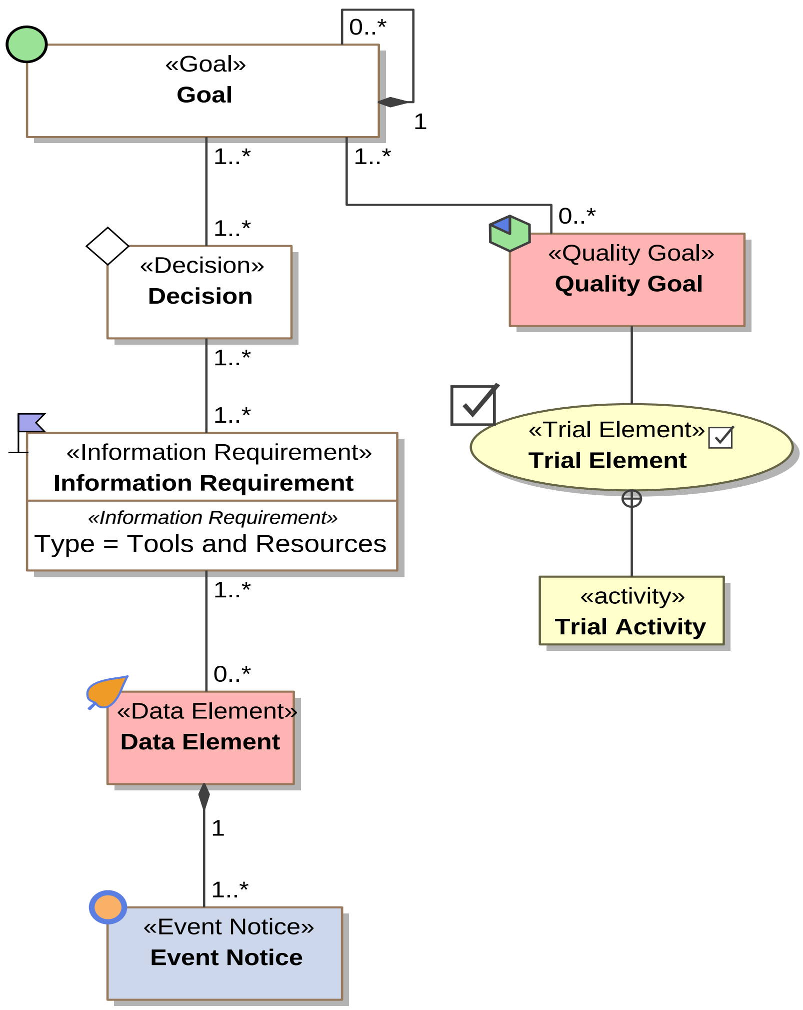


Figure 26: Intent Analysis Meta Model

The first step is to understand the domain through a cognitive engineering analysis. We recommend the Goal Directed Task Analysis (GDTA) which explicitly models the functional goals describing “why” and “what” type goals. Further details on GDTA analysis are in Chapter IV, Model Development. From this foundational understanding of the domain we extend the GDTA for Operationalized Intent.

Through the same SME interviews for the GDTA we identify candidates for the intent model which is traced to the functional goals of the GDTA. The quality goals and constraints should follow the heuristics in Chapter III, Model. The intent model should be developed and defined in collaboration with AIA designers, operators, and facilitated by cognitive engineers. There is no recommended number of quality goals or execution constraints. Having less than five quality goals is likely to result in simple behavior modalities while greater than ten begins to stretch the working memory of an operator.

Once there is an intent model, scenarios, vignettes, and missions which have ontological influence on the priority of the quality goals should be identified. These are decomposed into elements with a specific situation or disturbance. The trial elements are further decomposed into trial activities. The study trials are built by arranging trial elements in a situation – disturbance – situation sequence. The trial activities are assessed to define operator training requirements to support the study.

In parallel the information requirements of the GDTA are analyzed against the system design to identify data elements observable in the system which compose the information requirements. From these data elements a sampling method must be defined. Some changes in data elements are continuous (e.g. aircraft position), others are more

stochastic (e.g. airspace restriction location). The logging of an event notice should carry some significant and cohesive meaning. This may entail translating or deriving some information from data elements to populate the event notice. The details of event notice definition are in Chapter V, Data Model Development.

With a defined intent model, designed trials, and means of logging event notices, a study can be conducted to collect data. The study should include training on the operators on the use of the synthetic task environment (STE), Operationalized Intent, and any trial activity training requirements identified earlier. During the trial, the operator should be queried for their intent model at situationally relevant points. The philosophical and practical details are found in Chapter IV, Studying Intent and RPA Intent Study.

Using the analysis methods laid out in Chapter IV, Results and Analysis, a data centric evaluation of the trials and the intent model provides useful feedback. Intent model issues and trial emphasis biases can be identified in quality goal rank and intent change analyses. The effectivity of the operator training programs can be evaluated from a cohesion analysis.

The collected event notices should be translated and labeled appropriately to the situation. This translation should enhance the informational value from the perspective of the operator as captured in the GDTA. For example, from Chapter V Translation and Labeling, positions can be endogenously referenced as a range and bearing from locations relevant to the operator, in our case the aircraft position and the SPI. This supports the estimation algorithm identifying effective patterns instead of spurious ones.

From the ontology discussed earlier, intent estimation is primarily the ranking of quality goals. To accomplish this step, a label ranking algorithm is appropriate, as

discussed in Chapter V Estimation Algorithm. Any means of label ranking has strengths and weaknesses which must be assessed based on the situation data and implemented usage.

With an algorithm an estimation study can be executed using the situation data translated earlier. The estimates should be compared to the specific operator's elicited intent for that situation. This provides results for evaluating the accuracy, stability, sensitivity, and performance of the estimator. The details of these evaluations are illustrated in Chapter V, Results.

The information regarding the effectivity of the algorithm should be used to evaluate the event notice definitions or their translation into situation data. Identifying which data are least useful can reduce the overall system complexity. Conversely ontologically relevant data that lack utility to the estimator may suffer from translation or representation issues.

Once there is confidence in the intent model and the estimator, Operationalized Intent can be implemented in the target HAT. This involves implementing the intent model interpretation into the functional AIAs, integrating Operationalized Intent into the mental models taught in operator training, and developing a complete intent agent with an operator facing interface. These architectural and design details are beyond the scope of this research.

Contribution Summary

This research sought to discover new methods of improving coordination in HATs to enhance their team effectiveness. Through a thorough synthesis of coordination and intent, we have furthered the understanding and characterization of this critical area.

From the synthesis, the theory of Operationalized Intent provides a scalable framework for supporting intent aware multi-agent systems. This research produced a rigorous method to apply Operationalized Intent to a domain and study it effectively via human-in-the-loop testing. We demonstrated that the method was effective and able to characterize intent dynamics, operator cohesion, and provide feedback on the quality of the intent model and study trials. Finally, we demonstrated that it is possible to estimate dynamic intent with accuracy better than the situation baseline. This research successfully demonstrated the potential of Operationalized Intent as a new tool to improve coordination in human agent teams.

Future Work

The near term, next steps for this research are to address the estimation problem with improved tools. The recently released TensorFlow-Ranking library provides label ranking by applying deep learning. An analysis of the predictive utility of the situation vector features and the data elements would also lead to potentially improved estimation.

Beyond that, an expanded study with an iterated quality goal model and more intensive operator intent training is warranted to validate our initial findings. Such a study would provide a corpus of data which could identify more role-based trends.

Ultimately, studies of multi-operator, multi-AIA teams executing with and without the full intent models, quality goal hierarchy and execution constraint list, will validate the utility of intent aware AIAs. The final piece is developing intent projection capabilities integrated with coordinating plans to enable AIAs which anticipate their operator’s future needs.

Bibliography

Ahmad, B. I., Murphy, J. K., Langdon, P. M., Godsill, S. J., Hardy, R., & Skrypchuk, L. (2016). Intent Inference for Hand Pointing Gesture-Based Interactions in Vehicles. *IEEE Transactions on Cybernetics*, *46*(4), 878–889. https://doi.org/10.1109/TCYB.2015.2417053

Albert, W., & Tullis, T. (2013). *Measuring the user experience: collecting, analyzing, and presenting usability metrics*.

Andes, R. C. (1997). Assuring Human-Centeredness in Intelligent Rotorcraft Cockpits: Using Crew Intent Estimation to Coordinate RPA Functions. In *AHS International Forum*. AHS International.

Ashby, W. R. (1956). *An Introduction to Cybernetics* (2nd ed.). London: Chapman & Hall ltd. https://doi.org/10.2307/3006723

Banks, S. B., & Lizza, C. S. (1991). Pilot's Associate: a cooperative, knowledge-based system application. *IEEE Expert*, *6*(3), 18–29. https://doi.org/10.1109/64.87681

Bindewald, J. M., Peterson, G. L., & Miller, M. E. (2017). Clustering-based online player modeling. *Communications in Computer and Information Science*, *705*, 86–100. https://doi.org/10.1007/978-3-319-57969-6_7

Blanco, J. M., Illarramendi, A., & Goni, A. (1994). Building a Federated Relational Database System: an Approach Using a Knowledge-Based System. *International Journal of Cooperative Information Systems*, *03*(04), 415–455. https://doi.org/10.1142/S0218215794000211

Bolici, F., Howison, J., & Crowston, K. G. (2016). Stigmergic coordination in FLOSS development teams: Integrating explicit and implicit mechanisms. *Cognitive Systems Research*, *38*, 14–22. https://doi.org/10.1016/j.cogsys.2015.12.003

Bolstad, C. A., Riley, J. M., Jones, D. G., & Endsley, M. R. (2002). Using Goal Directed Task Analysis with Army Brigade Officer Teams. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *46*(3), 472–476. https://doi.org/10.1177/154193120204600354

Bonchek-Dokow, E., & Kaminka, G. A. (2014). Towards computational models of intention detection and intention prediction. *Cognitive Systems Research*, *28*, 44–79.

Boos, M., Kolbe, M., Kappeler, P. M., & Ellwart, T. (2011). Dimensions of Group Coordination: Applicability Test of the Coordination Mechanism Circumplex Model. In *Coordination in Human and Primate Groups* (pp. 57–73). https://doi.org/10.1007/978-3-642-15355-6

Bratman, M. E. (1990). Chapter 2: What is Intention? *Intentions in Communication*, 508.

Brooks, R. A. (1986). A Robust Layered Control System For A Mobile Robot. *IEEE Journal on Robotics and Automation*, *2*(1), 14–23. https://doi.org/10.1109/JRA.1986.1087032

Brundage, M. (2016). Modeling Progress in AI Analysis of Relevant Literatures. *The Workshops of the Thirtieth AAAI Conference on Artificial Intelligence*, 71–79.

Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. *Science*, *358*(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

Burke, C. S., Salas, E., Wilson-Donnelly, K., & Priest, H. (2004). How to turn a team of experts into an expert medical team: Guidance from the aviation and military

communities. *Quality and Safety in Health Care*, *13*(SUPPL. 1), 96–104. https://doi.org/10.1136/qshc.2004.009829

Butchibabu, A., Sparano-Huiban, C., Sonenberg, L., & Shah, J. (2016). Implicit Coordination Strategies for Effective Team Communication. *Human Factors*, *58*(4), 595–610. https://doi.org/10.1177/0018720816639712

Card, S. K., Moran, T. P., & Newell, A. (1983). *The Psychology of Human-Computer Interaction*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Carpenter, T. J., & Zachary, W. W. (2017). Using context and robot-human communication to resolve unexpected situational conflicts. *2017 IEEE Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2017*. https://doi.org/10.1109/COGSIMA.2017.7929596

Chen, J. Y. C., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical Issues in Ergonomics Science*, *19*(3), 259–282. https://doi.org/10.1080/1463922X.2017.1315750

Chen, M., Eccles, D. W., Hagen, H., Robert, S., Shneiderman, B., Tory, M., … Silver, D. (2009). Data, Information, and Knowledge in Visualization. *IEEE Computer Graphics and Applications*, *29*(1), 12–19. https://doi.org/10.1109/MCG.2009.6

Chen, Y. N., Sun, M., Rudnicky, A. I., & Gershman, A. (2016). Unsupervised user intent modeling by feature-enriched matrix factorization. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, *2016-May*, 6150–6154. https://doi.org/10.1109/ICASSP.2016.7472859

Cheruvelil, K. S., Soranno, P. A., Weathers, K. C., Hanson, P. C., Goring, S. J., Filstrup, C. T., & Read, E. K. (2014). Creating and maintaining high-performing collaborative research teams: The importance of diversity and interpersonal skills. *Frontiers in Ecology and the Environment*, *12*(1), 31–38. https://doi.org/10.1890/130001

Clark, H. H., & Brenan, S. E. (1991). Grounding in communication. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition* (pp. 127–149). Washington D.C.: American Psychological Association. https://doi.org/10.1201/b17574

Cooke, N. J., Gorman, J. C., Myers, C. W., & Duran, J. L. (2013). Interactive team cognition. *Cognitive Science*, *37*(2), 255–285. https://doi.org/10.1111/cogs.12009

Cox, A. (2017). *Functional Gain and Change Mechanisms in Post-Production Complex Systems*. https://doi.org/10.13140/RG.2.2.12946.76489

Cox, A. L., Cairns, P. A., Walton, A., & Lee, S. (2008). Tlk or txt? Using voice input for SMS composition. *Personal and Ubiquitous Computing*, *12*(8), 567–588. https://doi.org/10.1007/s00779-007-0178-8

Cox, A., & Szajnfarber, Z. (2018). System User Pathways to Change. In A. M. Madni, B. Boehm, R. G. Ghanem, D. Erwin, & M. J. Wheaton (Eds.), *Disciplinary Convergence in Systems Engineering Research* (pp. 617–634). Cham: Springer International Publishing.

Crowston, K. G. (1991). *Towards a Coordination Cookbook:Recipes for Multi-Agent Action*. *Alfred P. Sloan School of Management*. Massachusetts Institute of Technology.

Cuppens, F., Autrel, F., Miege, A., & Benferhat, S. (2002). Recognizing Malicious Intention in an Intrusion Detection Process. In *Second International Conference on Hybrid Intelligent Systems*. Santiago.

DeChurch, L. A., & Mesmer-Magnus, J. R. (2010). The Cognitive Underpinnings of Effective Teamwork: A Meta-Analysis. *Journal of Applied Psychology*, *95*(1), 32–53. https://doi.org/10.1037/a0017328

Delise, L. A., Gorman, C. A., Brooks, A. M., Rentsch, J. R., & Steele-Johnson, D. (2010). The Effects of Team Training on Team Outcomes: A Meta-Analysis. *Performance Improvement*, *22*(4), 53–80. https://doi.org/10.1002/piq

Diaconis, P., & Graham, R. L. (1977). Spearman's Footrule as a Measure of Disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*. Royal Statistical Society. https://doi.org/10.2307/2984804

Dickson, E. W., Singh, S., Cheung, D. S., Wyatt, C. C., & Nugent, A. S. (2009). Application of Lean Manufacturing Techniques in the Emergency Department. *The Journal of Emergency Medicine*, *37*(2), 177–182. https://doi.org/10.1016/J.JEMERMED.2007.11.108

Driskell, T., Salas, E., & Driskell, J. E. (2018). Teams in extreme environments: Alterations in team development and teamwork. *Human Resource Management Review*, *28*(4), 434–449. https://doi.org/10.1016/j.hrmr.2017.01.002

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank Aggregation Methods on the Web. *Proc. 10th World Wide Web Conf.*, 613–622. https://doi.org/10.1145/371920.372165

Eccles, D. W., & Tenenbaum, G. (2004). Why an expert team is more than a team of experts: A social-cognitive conceptualization of team coordination and communication in sport. *Journal of Sport and Exercise Psychology*, *26*(4), 542–560. https://doi.org/10.1123/jsep.26.4.542

Eccles, D. W., & Tenenbaum, G. (2007). A social-cognitive perspective on team functioning in sport. In G. Tenenbaum & R. C. Eklund (Eds.), *Handbook of sport psychology* (pp. 264–283). Hoboken, NJ US: John Wiley & Sons.

Endsley, M. R. (1988). Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 32, pp. 96–96). https://doi.org/10.1177/154193128803200220

Endsley, M. R. (2000). Direct measurement of situation awareness: validity and use of SAGAT. In D. J. Garland & M. R. Endsley (Eds.), *Situation awareness analysis and measurement*. Mahwah, NJ: Lawrence Erlbaum Associates Publishers.

Endsley, M. R., & Jones, D. G. (2012). *Designing for Situation Awareness: An Approach to User-Centered Design*. Boca Raton, FL: CRC Press, Inc.

Endsley, M. R., & Kiris, E. O. (1995). The Out-of-the-Loop Performance Problem and Level of Control in Automation. *Human Factors*, *37*(2), 381–394. https://doi.org/10.1518/001872095779064555

Entin, E. E., & Serfaty, D. (1999). Adaptive Team Coordination. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*(2), 312–325. https://doi.org/10.1518/001872099779591196

Espinosa, A., Lerch, J., & Kraut, R. (2004). Explicit vs. Implicit Coordination Mechanisms and Task Dependencies: One Size Does Not Fit All. In E. Salas & S.

M. Fiore (Eds.), *Team cognition: understanding the factors that drive process and performance* (1st ed., pp. 107–129). Washington, D.C.: American Psychological Association.

Feng, L., & Upenn, I. L. (2015). Design and Verification for Autonomous Systems Interacting with Human Operators Human-Automation Interaction, (June).

Ferguson, G., & Allen, J. F. (2007). Mixed-Initiative Systems for Collaborative Problem Solving. *AI Magazine*, *28*(2), 23–32.

Fincannon, T. D., Keebler, J. R., Jentsch, F., Phillips, E., & Evans, A. W. (2011). Team Size, Team Role, Communication Modality, and Team Coordination in the Distributed Operation of Multiple Heterogeneous Unmanned Vehicles. *Journal of Cognitive Engineering and Decision Making*, *5*(1), 106–131. https://doi.org/10.1177/1555343411399073

Fiore, S. M., & Wiltshire, T. J. (2016). Technology as teammate: Examining the role of external cognition in support of team cognitive processes. *Frontiers in Psychology*, *7*(OCT), 1–17. https://doi.org/10.3389/fpsyg.2016.01531

Franklin, S., & Graesser, A. (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents, 21–35. https://doi.org/10.1007/BFb0013570

Geddes, N. D. (1989). *Understanding Human Operators' Intentions in Complex Systems*. Georgia Institute of Technology.

Geddes, N. D. (1997). Associate Systems : A framework for human-machine cooperation. *Advances in Human Factors/Ergonomics*, 237–242.

Geddes, N. D., & Buchler, N. (2012). Warfighter Associate : Decision Aiding and Metrics for Mission Command.

Geddes, N. D., & Lizza, C. S. (2001). Practical Applications of a Real Time , Dynamic Model of Intentions.

Gervits, F., Thurston, D., Fong, T., Field, M., Pham, Q., & Scheutz, M. (2020). Toward Genuine Robot Teammates : Improving Human-Robot Team Performance Using Robot Shared Mental Models. *Proceedings of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020).*

Goertzel, B. (2014). Artificial General Intelligence: Concept, State of the Art, and Future Prospects. *Journal of Artificial General Intelligence*, *5*(1), 1–48. https://doi.org/10.2478/jagi-2014-0001

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge, MA: The MIT Press.

Gorman, J. C., Cooke, N. J., & Winner, J. L. (2006). Measuring team situation awareness in decentralized command and control environments. *Ergonomics*, *49*(12–13), 1312–1325. https://doi.org/10.1080/00140130600612788

Gorman, Jamie C., Amazeen, P. G., & Cooke, N. J. (2010). Team Coordination Dynamics. *Nonlinear Dynamics Psychology and Life Sciences*, *14*(3), 265–289. https://doi.org/10.4319/lo.2013.58.2.0489

Gorman, Jamie C., Cooke, N. J., Pederson, H. K., & Connor, O. O. (2005). Coordinated Awareness of Situation by Teams (CAST): Measuring team situation awareness of a communication glitch. In *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (pp. 808–812).

Hare, N., & Coghill, P. (2016). The future of the intelligence analysis task. *Intelligence*

and National Security, *31*(6), 858–870. https://doi.org/10.1080/02684527.2015.1115238
Hollnagel, E., & Woods, D. D. (2005). *Joint Cognitive Systems: Foundations of Cognitive Systems Engineering*. Boca Raton, FL: CRC Press.
Holtzen, S., Zhao, Y., Gao, T., Tenenbaum, J. B., & Zhu, S. C. (2016). Inferring human intent from video by sampling hierarchical plans. *IEEE International Conference on Intelligent Robots and Systems*, *2016-Novem*(1), 1489–1496. https://doi.org/10.1109/IROS.2016.7759242
Howard, N., & Cambria, E. (2013). Intention awareness: improving upon situation awareness in human-centric environments. *Human-Centric Computing and Information Sciences*, *3*(1), 9. https://doi.org/10.1186/2192-1962-3-9
Huang, R., Liang, H., Zhao, P., Yu, B., & Geng, X. (2017). Intent-Estimation- and Motion-Model-Based Collision Avoidance Method for Autonomous Vehicles in Urban Environments. *Applied Sciences*, *7*(5), 457. https://doi.org/10.3390/app7050457
Huber, C., & Marvel, L. M. (2016). Cyber Fighter Associate.
Hubert, T., Schrittwieser, J., Baker, L., Hui, F., Hassabis, D., Antonoglou, I., … Guez, A. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359. https://doi.org/10.1038/nature24270
Huget, M. P. (2014). Agent Communication. In *Agent-Oriented Software Engineering* (pp. 101–133). Berlin, Heidelberg: Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-54432-3_6
Humphrey, C. M., & Adams, J. A. (2011). Analysis of complex team-based systems: Augmentations to goal-directed task analysis and cognitive work analysis. *Theoretical Issues in Ergonomics Science*, *12*(2), 149–175. https://doi.org/10.1080/14639221003602473
Huntsberger, T. (2011). Cognitive architecture for mixed human-machine team interactions for space exploration. *IEEE Aerospace Conference Proceedings*, 1–11. https://doi.org/10.1109/AERO.2011.5747552
Jacques, D. R., & Strouble, D. D. (2010). *A-10 Thunderbolt II (Warthog) Systems Engineering Case Study*. Dayton, Ohio.
Jeffery, A. B., Maes, J. D., & Bratton-Jeffery, M. F. (2005). Improving team decision-making performance with collaborative modeling. *Team Performance Management*, *11*(1–2), 40–50. https://doi.org/10.1108/13527590510584311
Johnson-Laird, P. N. (1980). Mental Models in Cognitive Science. *Cognitive Science*, *115*(1), 71–115. https://doi.org/10.3354/meps09814
Johnson, C., Miller, M. E., Rusnock, C. F., & Jacques, D. R. (2020). Applying Control Abstraction to the Design of Human – Agent Teams. *Systems*. Retrieved from https://www.mdpi.com/2079-8954/8/2/10
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
Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C., Van Riemsdijk, B., & Sierhuis, M. (2012). Autonomy and interdependence in human-agent-robot teams.

IEEE Intelligent Systems, *27*(2), 43–51. https://doi.org/10.1109/MIS.2012.1
Johnson, M., Vignati, M., & Duran, D. (2019). Understanding Human-Autonomy Teaming through Interdependence Analysis. In J. Kerstholt, J. Barnhoorn, T. Hueting, & L. Schuilenborg (Eds.), *Automation as an Intelligent Teammate: Social Psychological Implications* (pp. 1–20). North Atlantic Treaty Organization Science and Technology Organization.
Jones, N. a., Ross, H., Lynam, T., Perez, P., & Leitch, A. (2011). Mental Model an Interdisciplinary Synthesis of Theory and Methods. *Ecology and Society*, *16*(1), 46–46. https://doi.org/46
Kaber, D. B., & Endsley, M. R. (2007). The effects of level of automation and adaptive automation on human performance, situation awareness and workload in a dynamic control task. *Theoretical Issues in Ergonomics Science*, *5*(2), 113–153. https://doi.org/10.1080/1463922021000054335
Kaber, D. B., Riley, J. M., Tan, K., & Endsley, M. R. (2001). On the Design of Adaptive Automation for Complex Systems. *International Journal of Cognitive Ergonomics*, *5*(1), 37–57.
Kearney, E., Gebert, D., & Voelpel, S. C. (2009). When And How Diversity Benefits Teams: The Importance Of Team Members' Need For Cognition. *Academy of Management Journal*, *52*(3), 581–598. https://doi.org/10.5465/amj.2009.41331431
Kelley, R., King, C., Tavakkoli, A., Nicolescu, M. M., Nicolescu, M. M., & Bebis, G. (2008). An Architecture for Understanding Intent Using a Novel Hidden Markov Formulation. *International Journal of Humanoid Robotics*, *5*(02), 203. https://doi.org/10.1142/S0219843608001418
Klein, G. (2008). Naturalistic decision making. *Human Factors*, *50*(3), 456–460. https://doi.org/10.1518/001872008X288385
Klein, G., & Bradshaw, J. M. (2005). Common Ground and Coordination in Joint Activity. In W. B. Rouse & K. R. Boff (Eds.), *Organizational Simulation* (pp. 139–184). Hoboken, NJ: Wiley-Interscience.
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*, *19*(6), 91–95. https://doi.org/10.1109/MIS.2004.74
Klinger, D., & Klein, G. (1999). An "Accident" Waiting to Happen? *Ergonomics in Design: The Quarterly of Human Factors Applications*, *7*(3), 20–25. https://doi.org/10.1177/106480469300100312
Kofler, C., Bhattacharya, S., Larson, M., Chen, T., Hanjalic, A., & Chang, S. F. (2015). Uploader Intent for Online Video: Typology, Inference, and Applications. *IEEE Transactions on Multimedia*, *17*(8), 1200–1212. https://doi.org/10.1109/TMM.2015.2445573
Kolbe, M., Burtscher, M. J., & Manser, T. (2013). Co-ACT-A framework for observing coordination behaviour in acute care teams. *BMJ Quality and Safety*, *22*(7), 596–605. https://doi.org/10.1136/bmjqs-2012-001319
Krozel, J., & Andrisani, D. (2008). Intent Inference with Path Prediction. *Journal of Guidance, Control, and Dynamics*, *29*(2), 225–236. https://doi.org/10.2514/1.14348
Kurzweil, R. (2005). *The singularity is near : when humans transcend biology*. Viking.
Lacerenza, C. N., Salas, E., Burke, C. S., Marlow, S. L., & Paoletti, J. (2017). Does team

communication represent a one-size-fits-all approach?: A meta-analysis of team communication and performance. *Organizational Behavior and Human Decision Processes*, *144*(February 2016), 145–170. https://doi.org/10.1016/j.obhdp.2017.08.001

Lagan-Fox, J., Anglim, J., & Wilson, J. R. (2004). Mental Models, Team Mental Models, and Performance: Process, Development and Future Directions. *Human Factors and Ergonomics in Manufacturing*, *14*(4), 331–352. https://doi.org/10.1002/hfm.20004 Mental

Lai, H. Y., Chen, C. H., Zheng, P., & Khoo, L. P. (2020). Investigating the evolving context of an unstable approach in aviation from mental model disconnects with an agent-based model. *Reliability Engineering and System Safety*, *193*(September). https://doi.org/10.1016/j.ress.2019.106657

Langan-Fox, J., Code, S., & Langfield-Smith, K. (2000). Team Mental Models: Techniques, Methods, and Analytic Approaches. *Human Factors*, *42*(2), 242–271. https://doi.org/10.1518/001872000779656534

Lee, S., Hwang, I., & Leiden, K. (2015). Intent Inference-Based Flight-Deck Human-Automation Mode-Confusion Detection. *Journal of Aerospace Information Systems*, *12*(8), 503–518. https://doi.org/10.2514/1.i010331

Licklider, J. C. R. (1960). Man-Computer Symbiosis. *IRE Transactions on Human Factors in Electronics*, *HFE-1*(1), 4–11. https://doi.org/10.1109/THFE2.1960.4503259

Lind, M. (2003). Making sense of the abstraction hierarchy in the power plant domain. *Cognition, Technology & Work*, *5*(2), 67–81. https://doi.org/10.1007/s10111-002-0109-4

Malone, T. W., & Crowston, K. G. (1994). The Interdisciplinary Study of Coordination. *ACM Computing Surveys*, *26*(1), 87–119. https://doi.org/10.1145/174666.174668

Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). Remember Me A Temporally Based Framework and Taxonomy of Team Processes. *Academy of Management Review*, *26*(3), 356–376.

Matheus, C. J., Kokar, M. M., & Baclawski, K. (2003). A Core Ontology for Situation Awareness. In *Proceedings of Sixth International Conference on Information Fusion* (pp. 545–552). Cairns, Australia. Retrieved from http://www.ccs.neu.edu/home/kenb/pub/2003/26/public.pdf

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), 273–283. https://doi.org/10.1037/0021-9010.85.2.273

McComb, S. A., Green, S. G., & Compton, W. D. (1999). Project Goals, Team Performance, and Shared Understanding. *Engineering Management Journal*, *11*(3), 7–12. Retrieved from https://afit.idm.oclc.org/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eoah&AN=35882417&site=pfi-live

McGhan, C. L. R., Nasir, A., & Atkins, E. M. (2015). Human Intent Prediction Using Markov Decision Processes. *Journal of Aerospace Information Systems*, *12*(5), 393–397. https://doi.org/10.2514/1.i010090

Meltzoff, A. N. (2005). Imitation and other minds: The “like me” hypothesis.

Perspectives on Imitation: From Neuroscience to …, *2*(1), 55–77.
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: The Journal of the Human Factors and Ergonomics Society*, *58*(3), 401–415. https://doi.org/10.1177/0018720815621206
Mesmer-Magnus, J., Niler, A. A., Plummer, G., Larson, L. E., & DeChurch, L. A. (2017). The cognitive underpinnings of effective teamwork: a continuation. *Career Development International*, *22*(5), 507–519. https://doi.org/10.1108/CDI-08-2017-0140
Miller, C. A. (2017). The firefox fallacy: Why intent should be an explicit part of the external world in human automation interaction. In *Cognitive Systems Engineering: The Future for a Changing World* (pp. 269–294). https://doi.org/10.1201/9781315572529
Miller, C. A., & Hannen, M. D. (1998). User Acceptance of an Intelligent User Interface: A Rotorcraft Pilot's Associate Example. *A Rotorcraft Pilot's Associate Example*, 109–116. https://doi.org/http://doi.acm.org/10.1145/291080.291100
Miller, C. A., & Hannen, M. D. (1999). The Rotorcraft Pilot's Associate: design and evaluation of an intelligent user interface for cockpit information management. *Knowledge-Based Systems*, *12*(8), 443–456. https://doi.org/10.1016/S0950-7051(99)00035-0
Mohammed, S., Ferzandi, L., & Hamilton, K. (2010). Metaphor no more: A 15-year review of the team mental model construct. *Journal of Management*, *36*(4), 876–910. https://doi.org/10.1177/0149206309356804
Naikar, N. (2017). Cognitive work analysis: An influential legacy extending beyond human factors and engineering. *Applied Ergonomics*, *59*. https://doi.org/10.1016/j.apergo.2016.06.001
Noe, R. A., Mcconnell Dachner, A., Saxton, B., & Keeton, K. E. (2011). *Team Training for Long-duration Missions in Isolated and Confined Environments: A Literature Review, an Operational Assessment, and Recommendations for Practice and Research*. *TM-2011-216162*. Houston, TX.
Noyes, J. M., & Starr, A. F. (2011). A Comparison of Speech Input and Touch Screen for Executing Checklists in an Avionics Application. *The International Journal of Aviation Psychology*, *17*(3), 299–315. https://doi.org/10.1080/10508410701462761
Okhuysen, G. A., & Bechky, B. A. (2009). Coordination in Organizations: An Integrative Perspective. *The Academy of Management Annals*, *3*(1), 463–502. https://doi.org/10.1080/19416520903047533
Pasumarthi, R. K., Bruch, S., Wang, X., Li, C., Bendersky, M., Najork, M., … Wolf, S. (2019). TF-ranking: Scalable tensorflow library for learning-to-rank. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2970–2978. https://doi.org/10.1145/3292500.3330677
Periverzov, F., & Ilieş, H. (2015). IDS: The intent driven selection method for natural user interfaces. *2015 IEEE Symposium on 3D User Interfaces*, *3DUI 2015 - Proceedings*, 121–128. https://doi.org/10.1109/3DUI.2015.7131736
Peterson, L. M., & Bailey, L. L. (2001). *Controller-to-Controller Communication and Coordination Taxonomy (C4T)*. Washington D.C. Retrieved from DOT/FAA/AM-

01/19
Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive Systems Engineering*. New York, NY: John Wiley & Sons.
Rico, R., Gibson, C. B., Sánchez-Manzanares, M., & Clark, M. A. (2019). Building team effectiveness through adaptation: Team knowledge and implicit and explicit coordination. *Organizational Psychology Review*, *9*(2–3), 71–98. https://doi.org/10.1177/2041386619869972
Riley, V. (1989). A General Model of Mixed-Initiative Human-Machine Systems. In *Proceedings of the Human Factors Society Annual Meeting* (Vol. 33, pp. 124–128). Denver, CO. https://doi.org/10.1177/154193128903300227
Rosenberg, N. (1982). Learning by Using. In *Inside the black box: Technology and economics* (1st ed., pp. 120–140). Cambridge UK: Cambridge University Press.
Rouse, W. B. (1991). *Design for Success: A Human Centered Approach To Designing Successful Products and Systems*. New York, New York, USA: John Wiley & Sons.
Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE Transactions on Systems, Man, and Cybernetics*, *22*(6), 1296–1308. https://doi.org/10.1109/21.199457
Rouse, W. B., Geddes, N. D., & Curry, R. E. (1987). An Architecture for Intelligent Interfaces: Outline of an Approach to Supporting Operators of Complex Systems. *Human—Computer Interaction*, *3*(2), 87–122. https://doi.org/10.1207/s15327051hci0302_1
Rowe, A., Liggett, K. K., & Davis, J. (2009). Vigilant spirit control station: A research test bed for multi-UAS supervisory control interfaces. In *Proceedings of the Fifteenth International Symposium …* (pp. 1–6). Wright State University CORE Scholar.
Salas, E., Cooke, N. J., & Rosen, M. A. (2008). On Teams, Teamwork, and Team Performance: Discoveries and Developments. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *50*(3), 540–547. https://doi.org/10.1518/001872008X288457
Salas, E., Dickinson, T., Converse, S., & Tannenbaum, S. (1992). Toward an understanding of team performance and training. Retrieved from http://psycnet.apa.org/record/1992-98450-001
Schaefer, K. E., Straub, E. R., Chen, J. Y. C., Putney, J., & Evans, A. W. (2017). Communicating intent to develop shared situation awareness and engender trust in human-agent teams. *Cognitive Systems Research*. https://doi.org/10.1016/j.cogsys.2017.02.002
Schaeffer, J. (2009). *One Jump Ahead*. *Chemical & Engineering News*. New York City, NY: Springer. https://doi.org/10.1021/cen-v030n028.p2919
Scheutz, M., DeLoach, S. A., & Adams, J. A. (2017). A Framework for Developing and Using Shared Mental Models in Human-Agent Teams. *Journal of Cognitive Engineering and Decision Making*, *11*(3), 203–224. https://doi.org/10.1177/1555343416682891
Schneider, M. F., & Miller, M. E. (2018). Operationalized Intent for Communication in Human-Agent Teams. *Proceedings - 2018 IEEE International Conference on Cognitive and Computational Aspects of Situation Management, CogSIMA 2018*,

117–123. https://doi.org/10.1109/COGSIMA.2018.8423992
Schneider, M. F., Miller, M. E., Jacques, D. R., Peterson, G. L., & Ford, T. C. (n.d.). Exploring the Impact of Coordination in Human-Agent Teams. *Cognitive Engineering And Decision Making*.
Schneider, M. F., Miller, M. E., Jacques, D. R., Peterson, G. L., & Ford, T. C. (2020). *Intent for Human-Agent Teaming*. Dayton, Ohio.
Schneider, M. F., Miller, M. E., & McGuirl, J. M. (2020). *A Study of Operationalized Intent for Remotely Piloted Aircraft*. Dayton, Ohio.
Searle, J. R. (1969). *Speech Acts: An Essay in the Philosophy of Language*. New York: Cambridge University Press.
Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
Shneiderman, B., Plaisant, C., Cohen, M., & Jacobs, S. (2010). *Designing the User Interface: Strategies for Effective Human-Computer Interaction* (5th ed.). TBS.
Stachowski, A. A., Kaplan, S. A., & Waller, M. J. (2009). The Benefits of Flexible Team Interaction During Crises. *Journal of Applied Psychology*, *94*(6), 1536–1543. https://doi.org/10.1037/a0016903
Stanton, N. A., Plant, K. L., Revell, K. M. A., Griffin, T. G. C., Moffat, S., & Stanton, M. (2019). Distributed cognition in aviation operations: a gate-to-gate study with implications for distributed crewing. *Ergonomics*, *62*(2), 138–155. https://doi.org/10.1080/00140139.2018.1520917
Sterling, L. S., & Taveter, K. (2009). *The Art of Agent-Oriented Modeling*. Boston, MA: MIT Press.
Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *41*(1), 61–71. https://doi.org/10.1518/001872099779577273
Stowers, K., Kasdaglis, N., Newton, O., Lakhmani, S., Wohleber, R., & Chen, J. Y. C. (2016). Intelligent agent transparency: The design and evaluation of an interface to facilitate human and intelligent agent collaboration. In *Proceedings of the Human Factors and Ergonomics Society* (Vol. 0). https://doi.org/10.1177/1541931213601392
Sukthankar, G., Shumaker, R., & Lewis, M. (2013). Intelligent agents as teammates. In *Theories of Team Cognition: Cross-Disciplinary Perspectives* (pp. 313–344). https://doi.org/10.4324/9780203813140-25
Sycara, K., & Lewis, M. (2004). Integrating Intelligent Agents into Human Teams. In E. Salas & S. M. Fiore (Eds.), *Team cognition: understanding the factors that drive process and performance* (1st ed., pp. 203–231). Washington, D.C.: American Psychological Association.
Tomai, E., & Forbus, K. D. (2009). EA NLU: Practical Language Understanding for Cognitive Modeling. In *Proceedings of the 22nd International Florida Artificial Intelligence Research Society Conference* (pp. 117–122). Sanibel Island, FL: Association for the Advancement of Artificial Intelligence.
van den Bossche, P., Gijselaers, W., Segers, M., Woltjer, G., & Kirschner, P. (2011). Team learning: Building shared mental models. *Instructional Science*, *39*(3).

https://doi.org/10.1007/s11251-010-9128-3
Vaniya, S., Lad, B., & Bhavsar, S. (2011). A Survey on Agent Communication Languages. *2011 International Conference on Innovation, Management and Service*, *14*, 237–242.
Vembu, S., & Gärtner, T. (2011). Label Ranking Algorithms: A Survey. In J. Fürnkranz & E. Hüllermeier (Eds.), *Preference Learning* (pp. 45–64). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg. https://doi.org/10.1007/978-3-642-14125-6
Vered, M., Kaminka, G. A., & Biham, S. (2016). Online Goal Recognition through Mirroring: Humans and Agents. *Advances in Cognitive Systems*, *4*.
Weiss, G. (2013). *Multiagent Systems*. Cambridge, MA: MIT Press.
Wohleber, R. W., Calhoun, G. L., Funke, G. J., Ruff, H., Chiu, C. Y. P., Lin, J., & Matthews, G. (2016). The impact of automation reliability and operator fatigue on performance and reliance. *Proceedings of the Human Factors and Ergonomics Society*, (September), 211–215. https://doi.org/10.1177/1541931213601047
Woods, D. D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering*. Boca Raton, FL, FL: CRC Press.
Xing, Y., Lv, C., Wang, H., Wang, H., Ai, Y., Cao, D., … Wang, F.-Y. (2019). Driver Lane Change Intention Inference for Intelligent Vehicles: Framework, Survey, and Challenges. *IEEE Transactions on Vehicular Technology*, *PP*(c), 1–1. https://doi.org/10.1109/TVT.2019.2903299
Yen, J., Fan, X., Sun, S., Hanratty, T., & Dumer, J. (2006). Agents with shared mental models for enhancing team decision makings. *Decision Support Systems*, *41*(3), 634–653. https://doi.org/10.1016/j.dss.2004.06.008
Zhou, Y., & Qiu, G. (2018). Random forest for label ranking. *Expert Systems with Applications*, *112*, 99–109. https://doi.org/10.1016/j.eswa.2018.06.036

Appendix A Glossary

Term	Definition
Human-Agent Team (HAT)	Group of humans and AIAs assembled to execute work towards a shared goal.
Artificial Intelligent Agent (AIA), agent	Systems capable of sensing their environment, reasoning about their situation, and taking action in response to changes in the environment. In this research they are the synthetic portion of the human-agent team. Functional AIAs and the Intent Agent are types of AIAs.
Intent	Relatively stable, pro-attitudes that function as inputs to further practical reasoning.
Goal Directed Task Analysis (GDTA)	Cognitive engineering analysis to capture the operator's goals in a domain, used in Situation Awareness design.
Situation	A term used in trial design to represent a stable state of the world, system, and team work.
Disturbance	A term applied during trial design to represent an event which alters a stable situation and forces a new situation to develop.
Quality Goal	A term originally defined by Sterling and Tavateer as a nonfunctional or quality requirement of a sociotechnical team. This term is applied here to represent a nonfunctional goal which can provide execution relevant guidance to AIAs regarding "how" they should execute their tasks.
Execution Constraint	A limiting condition on AIA behavior selection, typically to avoid violations of the performance envelope of the system controlled by the AIA.
Synthetic Task Environment (STE)	Experimental apparatus that mimics, to the maximum practical extent, the natural operational environment while still permitting a desirable level of experimental control.
Divergence	A distance metric for the dissimilarity of two ranked sets of the same items. Within this research, the metric is an exact normalized form of the Spearman Footrule.
Intent Cohesion	The extent to which a group of operators expressed similar quality goal rankings in similar situations. The distribution of divergence between pairwise rankings of operator elicitations in similar situations (all vs all).
Divergence Contribution	The extent to which a participant increases the total divergence of a population. Calculated, for a participant in a trial, based on a normalized ranked sum of operator median divergence across all elicitations.
Data Element	Observable data in a system that is pertinent to an information requirement of an operator.

Event Notice	A timestamped log of data demonstrating a a set of information relevant to a state of a subsystem, the environment, or other agents within the environment.
Situation Vector	Data structure representing the state of the world, system, and operator used to estimate intent. They are compiled and updated from event notice data which has been translated to an informational form.
Situation Baseline	The interquartile range of the divergence distribution from the operator responses within a situation, i.e. the intent cohesion.

Appendix B Data Cleaning Process

Initial Data Cleaning

Initial data cleaning of the elicitation data consisted of removing all training trials and participant 42 who did not complete the study. The timestamps were corrected from GMT to EST to synchronize with the events data. Detailed inspection indicated that Elicitation ID 57 was erroneously duplicative and removed.

For the event notice data, the training trials and participant 42 were also dropped from the dataset. There was a set of erroneous data recorded outside of any trial that was removed. Certain systems were renamed for consistency and all locations were translated from WGS84 to UTM42N coordinates.

Cohesion Analysis

As described in Chapter IV, Intent Cohesion, four trials were removed from the analysis. Participant 93 in trial 2 was removed because they suffered a simulator failure that meant their data could not be situationally correlated with the other participants throughout the entire trial. The cohesion analysis with all the remaining participants resulted in the cohesion being significantly above the top/bottom threshold and the Kruskal-Wallis between the participants indicated that their distributions were significantly different. Using a ranked sum method on the median divergence to develop a contribution score the three highest divergence contributors were removed from the three trials: Participant 93 from trial 1, Participant 57 from trial 2, and Participant 53 from trial 3. Having removed these participants, the Kruskal-Wallis test between the top/bottom threshold and the participants' divergence distribution was not significantly

different. The Kruskal-Wallis test between participants demonstrated that their distributions were not significantly different. It is worth noting that the contribution scores for the participants also changed, not just in value but in ordering.

Estimation Analysis

For the estimation analysis each elicitation was mapped to a specific situation to serve as a label for the situation vectors in that situation. For trial 1, all elicitations were used. In trial 2, elicitation 6 was not used for participants 53 and 39. These participants both had seven elicitations and the assignment of elicitation 6 would have overlapped with elicitation 4 in both cases. Observing the timelines, the events in proximity to the elicitations and the different between elicitation 4 and six indicated that it would be more appropriate to assigned elicitation 4 to the situation. This is consistent with the labeling of the other participants as well. In trial 3, elicitation 4 for participant 39 was not used. Elicitation 4 was mistakenly taken during the same situation as elicitation 3 and to be consistent with the labeling of the other participant's situations elicitation 3 was used. The effects of these exclusions or alternate mappings of elicitations to situation vectors were not explored in this research.

Appendix C Code Repository

Data Observer

Data observer is the Vigilant Spirit Control Station (VSCS) plug in that was used to extract data from VSCS and formulate it as event notices for the OI Study Database. It is written in C# using Microsoft Visual Studio 2017.

https://git.antcenter.net/hsi/dataobserver

Operationalized Intent

This repository contains the OI Study Database in PostgreSQL 11, the Python 3.7 analysis notebooks and scripts, and finally the OIE-RF estimator in MATLAB 2020a. It also includes all the SysML Models in Cameo Systems Modeler 19 and the Vigilant Spirit 6 file for the study.

https://git.antcenter.net/hsi/operationalized-intent

Appendix D IRB Package

IRB exception letter as embedded document.

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY (AETC)

HRPP Exempt Determination Form

For AFIT HRPP Use Only	
Protocol Number:	
Protocol Title:	

EDO Determination

Does this submission meet an Exempt Criteria? *Select the appropriate exemption category. Categories are defined in Exemption Request Package and on Page 2 of this form.*

☐ Yes	Which exempt category applies?	32 CFR 219.104 (d) (3) (i)	
	Is a limited IRB Review required to determine adequate provisions are in place to protect the privacy of subjects and maintain confidentiality of data?	☐ Yes	☐ No
	If a limited IRB review is required, IRB Member determined that *either*: ☐ Sufficient measures were taken to protect privacy and confidentiality. - OR - ☐ Insufficient measures were taken to protect privacy and confidentiality.		
☐ No	☐ The human subject research does not meet any exempt criteria. Referred to AFRL IRB Chair for IRB review.		
	- OR -		
	☐ The research uses an In Vitro diagnostic device with specimens that are NOT individually identifiable. Referred to AFRL IRB Chair to determine compliance with applicable FDA regulations.		

AFIT EDO / IRB Member Submission Analysis
EDO Reviewer Comments

AFIT EDO Signature	
	Click or tap to enter a date.
Exempt Determination Official	**Date**

Note: To sign this form electronically, please save it as a PDF and follow these instructions.

20 June 2019 1

Appendix E Additional Publications

These are additional publications not included in the body of the dissertation. The first two are related to this research and the rest are ancillary research efforts I supported during my PhD tenure.

Operationalized Intent for Communication in Human-Agent Teams

M. F. Schneider and M. E. Miller, "Operationalized Intent for Communication in Human-Agent Teams," Proc. - 2018 IEEE Int. Conf. Cogn. Comput. Asp. Situat. Manag. CogSIMA 2018, pp. 117–123, 2018.

COGSIMA
Operationalized Inten

Tracking Operator Intent in Tactical Operations

M. F. Schneider, M. E. Miller, and J. M. McGuirl, "Tracking Operator Intent in Tactical Operations," in IEEE International Conference on Systems, Man and Cybernetics., 2020.

SMC Tracking Intent
20200826 RevA.docx

Human Engagement with Event Rate Driven Adaptation of Automated Agents

M. F. Schneider, I. L. Bragg, J. P. Henderson, and M. E. Miller, "Human Engagement with Event Rate Driven Adaptation of Automated Agents," in IISE Annual Conference and Expo, 2018.

IISE paper Final
20180223.docx

Towards a meta-model to specify and design human-agent teams

M. F. Schneider, M. E. Miller, and J. M. Mcguirl, "Towards a meta-model to specify and design human-agent teams," in 20th International Symposium on Aviation Psychology, 2019.

ISAP Human-Agent
Modeling for Submiss

A SysML Language Extension and Method to Permit Modeling of Human-Agent Teams

M. Miller, J. McGuirl, M. F. Schneider, and T. Ford, "A SysML Language Extension and Method to Permit Modeling of Human-Agent Teams," Syst. Eng., 2020.

SEJ
SYS-19-140.R1_Proof_

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 074-0188*

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
14-09-2020	Dissertation	September 2017–September 2020

TITLE AND SUBTITLE	**5a. CONTRACT NUMBER**
Operationalized Intent for Improving Coordination in Human-Agent Teams	**5b. GRANT NUMBER**
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	**5d. PROJECT NUMBER**
Schneider, Michael, F, Civilian, USAF	**5e. TASK NUMBER**
	5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)	**8. PERFORMING ORGANIZATION REPORT NUMBER**
Air Force Institute of Technology Graduate School of Engineering and Management (AFIT/ENY) 2950 Hobson Way, Building 640 WPAFB OH 45433-8865	AFIT-ENV-DS-20-S-074
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**
US Air Force Office of Scientific Research Attn: Dr. James Lawton 4075 Wilson Blvd., Suite 350 Arlington, VA 2203 (703) 696-5999 (DSN: 426-5999), james.lawton.1@us.af.mil	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT
With the increasing capabilities of artificial intelligent agents (AIAs) integrated into multi-agent systems, future concepts include human-agent teams (HATs) in which the members perform fluidly as a coordinated team. Research on coordination mechanisms in HATs is largely focused on AIAs providing information to humans to coordinate better (i.e. coordination from the AIA to the human). We focus on the compliment where AIAs can understand the operator to better synchronize with the operator (i.e. from the human to the AIA). This research focuses specifically on AIA estimation of operator intent. We established the Operationalized Intent framework which captures intent in a manner relevant to operators and AIAs. The core of operationalized intent is a quality goal hierarchy and an execution constraint list. Designing a quality goal hierarchy entails understanding the domain, the operators, and the AIAs. By extending established cognitive systems engineering analyses we developed a method to define the quality goals and capture the situations that influence their prioritization. Through a synthesis of mental model evaluation techniques, we defined and executed a process for designing human studies of intent. This human-in-the-loop study produced a corpus of data which was demonstrated the feasibility of estimating operationalized intent.

15. SUBJECT TERMS
Human-Agent Teaming, Coordination, Intent Estimation, Operationalized Intent

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
					Michael Miller, AFIT/ENV
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(Include area code)*
U	U	U	UU	176	(937) 255-6565, ext 4651 (Michael.Miller@afit.edu)

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18