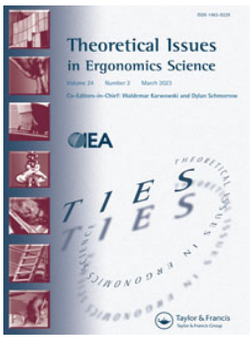

Theoretical Issues in Ergonomics Science

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ttie20

The role of shared mental models in human-AI teams: a theoretical review

Robert W. Andrews, J. Mason Lilly, Divya Srivastava & Karen M. Feigh

To cite this article: Robert W. Andrews, J. Mason Lilly, Divya Srivastava & Karen M. Feigh (2023) The role of shared mental models in human-AI teams: a theoretical review, Theoretical Issues in Ergonomics Science, 24:2, 129-175, DOI: 10.1080/1463922X.2022.2061080

To link to this article: https://doi.org/10.1080/1463922X.2022.2061080

Published online: 18 Apr 2022.

THEORETICAL ISSUES IN ERGONOMICS SCIENCE
2023, VOL. 24, NO. 2, 129–175
https://doi.org/10.1080/1463922X.2022.2061080

Taylor & Francis
Taylor & Francis Group

The role of shared mental models in human-AI teams: a theoretical review

Robert W. Andrews[a], J. Mason Lilly[b], Divya Srivastava[c] and Karen M. Feigh[a]

[a]School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, USA; [b]College of Computing, Georgia Institute of Technology, Atlanta, GA, USA; [c]School of Mechanical Engineering, Georgia Institute of Technology, Atlanta, GA, USA

ABSTRACT

Mental models are knowledge structures employed by humans to describe, explain, and predict the world around them. Shared Mental Models (SMMs) occur in teams whose members have similar mental models of their task and of the team itself. Research on human teaming has linked SMM quality to improved team performance. Applied understanding of SMMs should lead to improvements in human-AI teaming. Yet, it remains unclear how the SMM construct may differ in teams of human and AI agents, how and under what conditions such SMMs form, and how they should be quantified. This paper presents a review of SMMs and the associated literature, including their definition, measurement, and relation to other concepts. A synthesized conceptual model is proposed for the application of SMM literature to the human-AI setting. Several areas of AI research are identified and reviewed that are highly relevant to SMMs in human-AI teaming but which have not been discussed via a common vernacular. A summary of design considerations to support future experiments regarding Human-AI SMMs is presented. We find that while current research has made significant progress, a lack of consistency in terms and of effective means for measuring Human-AI SMMs currently impedes realization of the concept.

ARTICLE HISTORY
Received 12 November 2021
Accepted 29 March 2022

KEYWORDS
mental models; shared mental models; explainable AI; human-machine teaming; human-AI teaming

Relevance to human factors/ergonomics theory

This work summarizes the foundational research on the role of mental models and shared mental models in human teaming and explores the challenges of its application to teams of human and AI agents, which is currently an active topic of research both in the fields of psychology, human factors, cognitive engineering and human-robot interaction. A rich theoretical understanding herein will lead to improvements in the development of human-autonomy teaming technologies and collaboration protocols by allowing a critical aspect of teaming in the joint human-AI system to be understood in its historical context. The inclusion of a significant review of explainable AI will also enable researchers not associated with the development of AI systems to integrate and build upon the significant efforts of colleagues in HRI and AI more generally.

CONTACT Karen M. Feigh karen.feigh@gatech.edu Georgia Institute of Technology, Atlanta, GA, USA

1. Introduction

Mental model theory of an individual, initially put forward by Johnson-Laird in the early 1980s (Johnson-Laird 1983, 1980), is one attempt at describing an important aspect of human cognition: the way humans interpret and interact with an engineered environment. Mental models are generally defined as the abstract long-term knowledge structures humans employ to describe, explain, and predict the world around them (Converse, Cannon-Bowers, and Salas 1993; Van den Bossche et al. 2011; Johnson-Laird 1983; Rouse and Morris 1986; Scheutz, DeLoach, and Adams 2017; Norman 1987; Johnson-Laird 1980). In the 50 years since the term's inception, mental models have been the subject of significant scientific consideration and the basis for many theoretical and practical contributions to human-automation interaction, human-human teams, and human judgment and decision-making more generally.

Shared mental models (SMMs) are an extension of mental model theory, proposed by Converse, Salas, and Cannon-Bowers as a paradigm to study team training (Converse, Cannon-Bowers, and Salas 1993). The central idea of SMMs is that when individual team members' mental models align—when they have similar understandings of their shared task and each other's role in it—then this 'shared' mental model will allow the team to perform better because they will be able to more accurately predict the needs and behaviors of their teammates. The results of study on human-human SMMs have revealed that teams are indeed more effective when they are able to establish and maintain an SMM, see §2.

With the advent of advanced automation bordering on autonomy from advances in fields such as control theory, machine learning (ML), and artificial intelligence, human mental models are once again of interest as humans endeavor to create effective teams that incorporate automation with capabilities similar to those of a human teammate. However, unlike human-human teams, Human-AI teams[1] require differential study to fully capture the necessary bi-directional relationship—humans creating MMs of their team, and artificial agents creating MMs of their team— to fully create and understand Human-AI SMMs.

This paper presents a literature review of Human-AI SMMs culminating in a conceptual model. The paper fills the gap in the current literature on SMMs for Human-AI teams, which is currently disjointed, as researchers from diverse fields (psychology, cognitive science, robotics, human-robot interaction, cognitive engineering, human-computer interaction, etc.) investigate similar constructs using different terminology and methods and publish in distinct literatures. The paper seeks to lay the groundwork for future research on Human-AI SMMs by (1) reviewing and summarizing existing work on SMMs in humans, including their definitions, elicitation, measurement, and development; (2) defining Human-AI SMMs and situating them among the larger field of similar constructs; (3) identifying challenges facing future research; and (4) identifying new, developing, or previously unconsidered areas of research that may help to address these challenges. The scope of this review is focused on engineered socio-technical systems in which at least one human and one artificial agent work together in a team and seek to achieve common goals. We emphasize the definitions of the various constructs, conceptual and computational frameworks associated with SMMs, measurements and metrics used to measure SMMs, and open questions that remain.

This paper is divided into eight sections. In the next section, §2, we define and summarize mental models in the context of individual humans, including how they are elicited. The

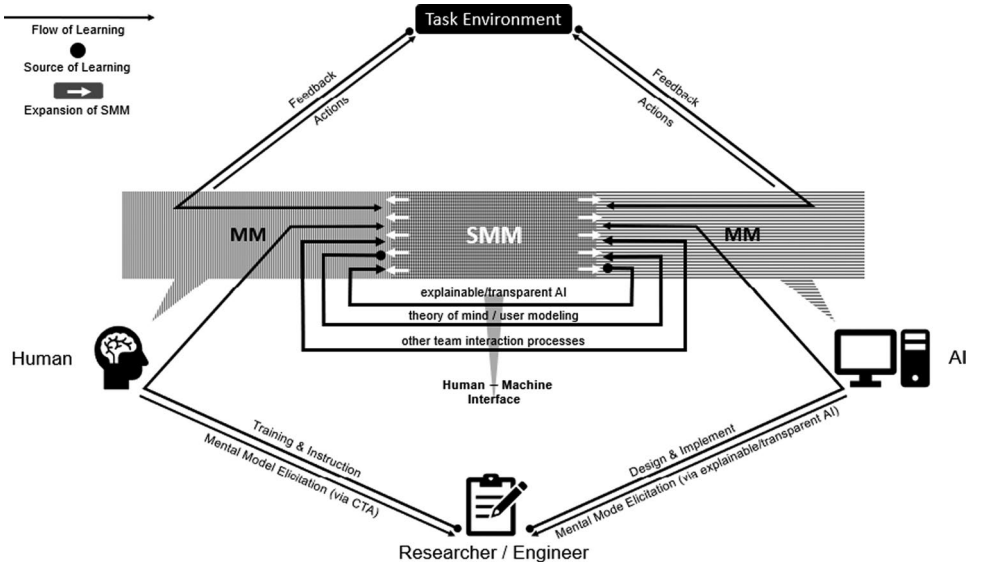

Figure 1. Diagram of an artificial SMM.

third section, §3, discusses SMMs for human-human teams, including metrics, research on their development, and related concepts from our literature review. The fourth section, §4, expands SMMs to human-AI teams. Here we present a conceptual model of Human-AI SMMs based on our synthesis of the literature (see Figure 1) and discuss the added difficulties of applying SMM theory to artificial agents. §5 covers some existing or emerging computational techniques for modeling human teammates, a key component of SMMs. §6 provides an overview of the explainable AI movement, which can be seen as intricately linked to the creation and maintenance of both human and artificial components of Human-AI SMMs. The penultimate section, §7, presents considerations for the design and study of Human-AI SMMs. We conclude in section §8.

2. Individual mental models & related concepts

Something like the idea of mental models had been around for decades before *mental model theory* was coined and popularized by Johnson-Laird in the 1980s. Kenneth Craik, for example, conceived of cognition as involving 'small-scale models' of reality as early as 1943 (Craik 1943). Jay W. Forrester, the father of system dynamics, speaks of 'mental images' and explains that 'a mental image is a model' and that 'all of our decisions are taken on the basis of models' (Forrester 1971). Forrester goes so far as to refer to these mental images as *mental models* and eloquently describes the concept as follows: 'One does not have a city or a government or a country in his head. He has only selected concepts and relationships which he uses to represent the real system.' Other authors referred to this idea as *internal representations* or *internal models* before Johnson-Laird (1980), and similar concepts going by different names have also been described and explored in fields such as psychology, computer science, military studies, manual and supervisory control, and learning science. In this section, we review the historical definition of mental models and situate them with similar concepts from neighboring fields and explain their relationship, where applicable.

We also discuss the subject of mental model elicitation, including methods and the difficulties involved.

Mental models are simplified system descriptions, employed by humans when systems, particularly modern socio-technical systems, are too complex for humans to understand in full detail. In the process of mental model creation, humans tend to maximize system simplification while minimizing forgone performance. Mental models are 'partial abstractions that rely on domain concepts and principles' (Mueller et al. 2019, 24). They enable understanding and control of systems at a fraction of the processing cost required by more comprehensive analytical strategies (Norman 1987).

Despite the goal of maintaining accuracy, some information is necessarily lost in the abstraction process. Norman explains that mental models often contain partial domain descriptions, workflow shortcuts, large areas of missing information, and vast amounts of uncertainty (Norman 1987). Many efforts to characterize expert mental models in system controls applications have 'resulted in hypothesized mental models that differ systematically from the "true" model of the system involved' (Rouse and Morris 1986; Rouse 1977; Van Bussel 1980; Van Heusden 1980; Jagacinski and Miller 1978). For example, humans often adhere to superstitious workflows that consistently produce positive results without fully grasping the system dynamics informing the utility of their strategies - Norman gives the example of users 'clearing' a calculator many more times than necessary. The knowledge structures underlying human cognition in these cases are practical, but far from accurate.

One way to enjoy the efficiencies of simplification while avoiding its consequences is to model the system at multiple levels of abstraction and use only the version with enough information for the task at hand. Rasmussen's work on cognitive task analysis (CTA) and work domain abstraction/decomposition suggests that humans engage in precisely such a process (Rasmussen 1979). A large repertoire of useful system representations at various levels of abstraction provides a solid basis from which system management strategies can be generated spontaneously, without needing to consider the whole of system complexity (Rasmussen 1979). These mental models allow domain practitioners to 'estimate the state of the system, develop and adopt control strategies, select proper control actions, determine whether or not actions led to desired results, and understand unexpected phenomena that occur as the task progresses' (Veldhuyzen and Stassen 1977).

Mental models are context-specific and/or context-sensitive (Converse, Cannon-Bowers, and Salas 1993; Scheutz, DeLoach, and Adams 2017; Rouse and Morris 1986). They are created to facilitate task completion in a given work domain—entirely different mental models are used in economic forecasting, for example, versus those employed in air-to-air combat.

Mental models are dynamic but slowly evolving, and they are generally considered to exist in long-term memory. As domain practitioners learn more about the work domain and experience a wider variety of off-normal operating conditions, mental models are updated to account for new information, and are thus also subject to considerable individual differences. How the content and structure of a mental model develop is actively being investigated, with some concluding that they develop gradually, with potential for greater rates of change occurring in less experienced operators (Endsley 2016), and others believing that they develop in 'spurts', only changing when a new model is presented to replace them (Mueller et al. 2019, 74; Einhorn and Hogarth 1986).

2.1. Mental model elicitation

Mental model elicitation is the process of capturing a person's mental model in a form that can examined by others. This can be used to understand a complex sociotechnical process, evaluate the design of a system, or provide insight about how mental models develop or change in certain conditions.

Formally, the mental model that is measured is not the same as the mental model that actually exists. Norman distinguishes between four things: (1) A system, (2) an engineer's model of the system (i.e., the design documents), (3) a user's mental model of the system, and (4) a researcher's mental model of the user's mental model (Norman 1987). The resulting representation - a model *of* the model - will take different forms depending on the methods used and the needs of the researcher, and it may be subjective or objective, qualitative or quantitative. The methods used to elicit mental models overlap strongly with the field of cognitive task analysis (Bisantz and Roth 2007) and can be subdivided into observational methods, interviews, surveys, process tracing, and conceptual methods (Cooke et al. 2000). These tactics may be used individually or in combination.

Observation consists simply of watching an operator's performance and thereby drawing conclusions about their reasoning. This could take a number of forms, including directly monitoring the operator in real time from a distance, embedding an undercover observer into the work domain, or filming the performance of a task (Bisantz and Roth 2007). Observational techniques offer intimate detail of how a task is performed; however, without adequate precautions, many methods risk interfering with normal operation. This approach is straightforward and can be useful, but it is sometimes criticized as subjective because of the potential biases of the observer and for producing largely qualitative information (Bisantz and Roth 2007; Cooke et al. 2000).

Interviews attempt to extract a mental model through conversation with an operator and can be classified as structured, unstructured, or semi-structured (Cooke et al. 2000; Bisantz and Roth 2007). Unstructured interviews are free-form and give the interviewer the opportunity to probe domain practitioners comprehensively on any areas of interest, such as asking for narrations of critical events or any idiosyncrasies of the person's performance. Structured interviews aim to be more rigorous or objective by presenting the same questions in the same order to multiple people, enabling a degree of quantifiability. Nonetheless, interviews are still susceptible to biases since they depend on verbal questions and responses, as well as the interviewee's introspection.

Surveys are quizzes that attempt to capture a mental model in a specific set of quantifiable questions. They have a broad variety of uses, including characterizing a practitioner's model of a task in general or investigating how a specific performance evolved. Common examples include Likert scales and subject-matter aptitude tests. Surveys can be sufficient in isolation, but they can also be used to supplement other elicitation methods such as observation. Surveys are closely related to structured interviews, only given asynchronously and with a finite selection of answers. In this way, they lose some expressive power but produce more quantifiable results—necessary for many applications. Both surveys and interviews can be used to quiz the operator's depth of understanding of the environment or to provide insight to how the operator thinks via introspection. Sarter and Woods note in particular the usefulness of mid-operation surveys for measuring *situation awareness*, a closely related concept to mental models (see below) (Sarter and Woods 1991). A specific example of this for

simulated environments is Endsley's Situation Awareness Global Assessment Technique (SAGAT), which occasionally halts a simulation to quiz the user about the state of the world (Endsley 1995).

Process tracing can be used to analyze user mental models by collecting and processing electronic data directly from the work domain, including user behavioral records, system state information, verbal reports from the user, and eye movements (Patrick and James 2004). These methods can be advantageous because they provide direct, quantitative information that is not subject to the biases of introspection (Cooke et al. 2000). An example of a process tracing success is Fan and Yen's development of a *Hidden Markov Model* for quantifying cognitive load based on secondary task performance (Fan and Yen 2011). However, in many cases, process tracing has not produced such conclusive results. The use of eye tracking in particular has proven difficult to directly correlate with the internal cognition that causes the observed motions.

Conceptual methods generate spatial representations of key concepts in the work domain, along with the constraints and 'shunts' that link them together in the mind of the practitioner (Cooke et al. 2000; Bisantz and Roth 2007). Examples include concept mapping, cluster analysis, multidimensional scaling, card sorting, and Pathfinder (Mohammed, Klimoski, and Rentsch 2000; Cooke et al. 2000). Conceptual analysis elicitation methods are popular for mental model elicitation, as mental models are often thought to exist in a pictorial or image-like form (Rouse and Morris 1986). Once a conceptual map of the work domain has been developed, it can be cross checked against a global reference or the concept map of a teammate. Conceptual methods are the most common among the methods described herein (Lim and Klein 2006; Mathieu et al. 2000; Van den Bossche et al. 2011; Marks, Zaccaro, and Mathieu 2000; Resick et al. 2010).

A central difficulty of measuring mental models is that the researcher can only measure what they think to ask. This means that unconventional, emergent behaviors are likely to be difficult or impossible to capture. It also requires the researcher themselves to be familiar with the system being modeled. (This familiarity might result from some kind of cognitive task analysis of the work domain (Naikar 2005).) Additionally, it must be noted that when measuring mental models, the mere act of asking questions may influence the models' future development by directing the subject's attention to previously unconsidered subjects or by restricting it to specific ideas proposed by the question. For example, 'how fast can the car accelerate?' suggests this is an important quantity to consider, whereas 'to what degree is the AI trustworthy?' may influence the operator to be more skeptical.

Mental model elicitation remains an inexact discipline. All available methods are vulnerable to various biases and confounding variables. French et al. succinctly discuss these difficulties in their review of trust in automation, drawing a connection between trust and mental models as 'hypothetical latent construct[s] which cannot be directly observed or measured but only inferred' (French, Duenser, and Heathcote 2018, 49-53). Ultimately, there are two ways to measure such a construct: by drawing inferences from observations of an operator's behavior, or by introspection on the part of the operator—and neither affords a direct view of the underlying construct. Methods relying on introspection, including surveys, interviews, and conceptual methods, provide some insight into the operator's thought processes; however, it is generally acknowledged that introspection is not fully reliable. Additionally, all such methods are limited by how the participant interprets the

questions posed to them. By contrast, behavioral methods, including observation and process tracing, eliminate the need to ask subjective questions but are confounded by the many other factors that may influence behavioral outcomes, including workload, stress, and fatigue.

2.2. Related concepts

Mental models' qualities place them alongside at least three other constructs: knowledge representation, schema, and situation awareness.

2.2.1. Knowledge representation

Knowledge Representation is broad term prevalent in both psychology and AI literature that concerns concrete models of the information stored in minds. In psychology (Rumelhart and Ortony 1977), the focus is on deriving these structures as naturally found in humans, and a subset of this work includes efforts to externalize people's mental models. AI researchers have also used this term in the context of producing 'knowledge-based AI' systems that mimic an information processing model of the human mind—systems that could be considered 'artificial mental models.'

2.2.2. Schema

Rumelhart and Ortony (1977) define *schema* as 'data structures for representing the generic concepts stored in memory,' which can represent objects, situations, events, actions, sequences, and relationships. Schema (sometimes pluralized 'schemata') are the subject of a vast body of literature in the psychology and cognitive science fields (McVee, Dunsmore, and Gavelek 2005; Stein 1992).

Four essential characteristics of schema are given (Rumelhart and Ortony 1977):

1. Schema have variables
2. Schema can embed one within the other
3. Schema represent generic concepts which, taken altogether, vary in their levels of abstraction
4. Schema represent knowledge rather than definitions

Schema has been discussed widely and their precise definition is at times ambiguous. Some authors view schema as functionally identical to mental models, while Converse et al. present schema as static data structures (as opposed to mental models, which both organize data and can be 'run' to make predictions) (Converse, Cannon-Bowers, and Salas 1993, 227). One clear difference is their domain of application: Mental model theory has primarily been developed in relation to engineered, socio-technical systems, whereas schema theory has been developed in more naturalistic domains: language, learning, social concepts, etc. The idea of schema predates mental models (Converse, Cannon-Bowers, and Salas 1993; Endsley 2016).

Scripts are a type of schema that organize procedural knowledge (Dionne et al. 2010). They are prototypical action sequences that allow activities to be chunked into hierarchically larger abstractions (Schank and Abelson 1977).

2.2.3. Situation awareness

Situation awareness (SA) refers to the completeness of one's knowledge of the facts present in a given situation. This state of knowledge arises from a combination of working memory, the ability to perceive information from the environment, and the mental models necessary to make accurate inferences from that information. The concept of SA, originally put forth by Mica Endsley, has been thoroughly studied from the perspective of air-to-air combat and military command and control. The general thrust of this subfield has been that a large proportion of accidents in critical systems can be attributed to a loss of situation awareness.

Endsley (Endsley 2004, 13-18) defines SA as existing at three levels, with awareness at each level being necessary for the formation of the next:

1. Level 1: Perception of Elements in the Environment. Deals with raw information inputs, their accuracy, and the ability of the human to access them and devote sufficient attention to each.
2. Level 2: Comprehension of the Current Situation. Refers to the human's ability to understand the implications of the raw inputs—the ability to recognize patterns and draw inferences.
3. Level 3: Projection of Future Status. Refers to the ability to anticipate how the system will evolve in the near future, given its current state.

What may blur the distinction between SA and mental models is Endsley's formal definition of SA as 'the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future' (Gilson et al. 1994; Wickens 2008; Endsley 2004; Endsley 2016). Taken on its own, this is very similar to the common definition of mental models as that which is used to 'describe, explain, and predict' the environment. The key difference is that mental models are something that exists in long-term memory, regardless of present context, whereas situation awareness arises within the context of a specific situation and results from the *application* of mental models to stimuli and working memory. Mental models aid situation awareness by providing expectations about the system, which guides attention, provides 'default' information, and allows comprehension and prediction to be done 'without straining working memory's capabilities' (Endsley 2016; National Academies of Sciences Engineering and Medicine 2022).

This same relationship underlies one of the key inconsistencies across all the following literature - the distinction between models and the states of those models - with some studies claiming to have ascertained one's mental model by deducing the current state of their short-term knowledge. We take the view in this paper that a *model* is, in all cases, both long-term and *functional* (it can be 'run') (see also Endsley 2004, 21). States, like situation awareness, are transient and informed by their underlying models. This close relationship makes measuring one useful for, but not equivalent to, measuring the other.

2.2.4. Summary

All of the preceding terms describe similar phenomena from different subfields of the human sciences. Mental model theory is a perspective drawing largely from seminal works in psychology, applied most commonly in the context of engineered socio-technical systems.

Although psychologists were initially interested in the implications of these knowledge structures in a single human, a newer branch of research has focused on the interaction of mental models in groups and work teams (i.e., SMMs). We describe these and related constructs in the next section.

3. Shared mental models (SMMs) & related concepts

SMMs are an extension of mental model theory, proposed and popularized by Converse, Salas, and Cannon-Bowers as a paradigm to study training practices for teams (Converse, Cannon-Bowers, and Salas 1993). The core hypothesis is that if team members have similar mental models of their shared task and of each other, then they are able to accurately predict their teammates' needs and behaviors. This facilitates anticipatory behavior and, in turn, increases team performance. Research has investigated how SMMs can be fostered through training, which models should be shared, and how much overlap is appropriate. It has also investigated how degrees of mental model sharing can be evaluated, to test the hypothesis that SMMs lead to improved performance.

3.1. Structure and dynamics of SMMs

A team is defined in this context as a group of individuals with differing capabilities and responsibilities working cooperatively toward a shared goal. Mental effort in a team is spent on at least two functions: interaction with the task environment, and coordination with the other team members, which Converse et al. term 'task work' and 'team work' (Converse, Cannon-Bowers, and Salas 1993). Since they deal with distinct systems (the task environment and the team), these functions use separate mental models. Converse et al. propose as many as four models per individual:

1. The Task Model — the task, procedures, possible outcomes, and how to handle them; 'what is to be done'
2. The Equipment Model — the technical systems involved in performing a task and how they work; 'how to do it'
3. The Team Model — the tendencies, beliefs, personalities, etc. of one's teammates; 'what are my teammates like'
4. The Team Interaction Model — the structure of the team, its roles, and the modes, patterns, and frequency of communication; 'what do we expect of each other'

Subsequent literature (e.g., Scheutz, DeLoach, and Adams 2017)) often reduces this taxonomy to two: the Task Model and Team Model.

An SMM is believed to benefit a team by '[enabling members] to form accurate explanations and expectations for the task, and, in turn, to coordinate their actions and adapt their behavior to demands of the task and other team members' (Converse, Cannon-Bowers, and Salas 1993, 228). The process by which SMMs form is not fully understood, though there is some consensus. Team members begin with highly variable prior mental models, which may depend on their past experiences, education, socio-economic background, or individual personality (Converse, Cannon-Bowers, and Salas 1993; Schank and Abelson 1977). Through training, practice with the task, and interaction with their teammates, they

gradually form mental models that enable accurate prediction of team and task behaviors. They may also adjust their behaviors to meet the expectations of their teammates. The exact interaction and learning processes involved, what training methods are most effective, and the rates at which SMMs develop and degrade are primary subjects of research on the topic. We detail some of the prominent works in this area below in §3.4.

3.2. Eliciting SMMs

The most common way to measure an SMM is to elicit mental models of team members individually and derive from these an overall metric of the team, as described in §3.3. However, other approaches examine the team as a whole, analyzing the emergent team knowledge structure as a distinct entity. Cooke argues that the latter approach may be the most important for fostering team performance but notes that very little research has done so (Cooke et al. 2004). Based on our analysis, this still seems to be the case, 18 years later.

The holistic approach advocated by Cooke calls for applying elicitation techniques to the team as a group (Cooke et al. 2004). For example, if questionnaires are employed, then the team would work together to select their collective responses; if conceptual methods are employed, then the team would cooperatively construct their domain representation. Such methods would help ensure that the same group dynamics that guide task completion (i.e., effects of team member personalities, leadership abilities, communication styles) are also present in SMM elicitation (Cooke et al. 2004; Dionne et al. 2010).

It is important to note that many experiments purporting to study SMMs do not elicit or apply any metric to them whatsoever. Rather, they may make a qualitative assertion of the presence of an SMM based on a specific intervention, such as cross-training, or some other structuring of the task environment. For example, one study equates the presence of shared displays with shared task models, and the reading of teammate job description with shared team models, but it does not explicitly attempt to measure the resultant or anticipated SMM (Bolstad and Endsley 1999).

The elicitation of MMs and the study of SMMs has predominantly been considered from the perspective of establishing a theoretical link between mental model sharedness and improved team performance. As of now, these elicitation methods are not being used to inform the conveyance or convergence of mental models in the human-human or human-AI domain.

3.3. Metrics for SMMs

Though ample work has been done on the elicitation of SMMs, relatively little literature exists on the key problem of synthesizing this information into salient metrics. Most authors agree that mental models are best measured in terms of the the *expectations they produce*, rather than their underlying representation because it is the expectations themselves that are thought to lead to improved performance (Converse, Cannon-Bowers, and Salas 1993; DeChurch and Mesmer-Magnus 2010; Jonker, Riemsdijk, and Vermeulen 2010; Cooke et al. 2000). Techniques to elicit these expectations run the gamut from casual interviews to structured surveys, but all literature on quantitative metrics assumes this information can be ultimately distilled to a discrete set of questions and answers. Examples of the questions involved might include 'Is team member A capable of X?', 'What is the maximum acceptable

speed of the vehicle?', 'When is it appropriate to send an email?', or 'Does team member B know Y?' Note that the questions need not have quantitative or bounded answers; they may be arbitrary sentences, provided that there is a sound way to test the equivalence of two answers (Jonker, Riemsdijk, and Vermeulen 2010).

Nancy Cooke's seminal work on measuring team knowledge from 2000 is still relevant today, and it has proven to provide the most valuable insight on this subject (Cooke et al. 2000). Her work suggests that once data have been elicited about each team member's mental model, or about the team holistically, at least six kinds of metrics can be used to relate SMMs to task performance: *similarity*, *accuracy*, *heterogeneous accuracy*, *inter-positional accuracy*, *knowledge distribution*, and *perceived mutual understanding* (Cooke et al. 2000; Burtscher and Oostlander 2019).

3.3.1. Similarity

Similarity refers to the extent to which team mental models are shared (i.e., are equivalent to each other). When using surveys for mental model elicitation, this may refer to the number or percentage of questions answered identically by a pair of teammates. For a concept map, it could refer to the number or percentage of links between domain concepts shared by a pair of teammates (Cooke et al. 2000).

It has been repeatedly shown that similarity alone in team mental models does lead to improved team performance: similarity facilitates shared expectations, which results in improved coordination. This appears to apply to both task models and team models. In a field study of 71 military combat teams, conceptual methods demonstrated that both task-work and teamwork similarity predicted team performance (Lim and Klein 2006). Conceptual methods, supported by some qualitative assertions, have been used to the same end in the domain of simulated air-to-air combat (Mathieu et al. 2000). Survey techniques and conceptual methods have also been used to show that similarity of task mental models is correlated with team effectiveness in the context of business simulations (Van den Bossche et al. 2011).

3.3.2. Accuracy

Cooke writes, 'all team members could have similar knowledge and they could all be wrong' (Cooke et al. 2000, 164). Accuracy measures aim to account for this by comparing each team member's task model to some 'correct' baseline. This is common with survey techniques, in which accuracy is the number or percentage of questions answered correctly. For conceptual methods, accuracy is the number or percentage of correctly identified domain concepts, as well as the links and constraints connecting them. Overall team accuracy is then measured as some average of these individual scores (Cooke et al. 2004).

Both qualitative and conceptual methods for team mental model elicitation have been used to show that accurate task models correlate with team effectiveness in military combat operations (Marks, Zaccaro, and Mathieu 2000; Lim and Klein 2006). Conceptual methods have also been used to demonstrate a positive correlation between mental model accuracy and team effectiveness in a simulated search and capture task domain (Resick et al. 2010). Though mental model accuracy does appear to play a role in teaming, some studies have found only a marginally significant effect (Webber et al. 2000).

Note that the correct baseline may sometimes be hard to identify. Jonker et al. identify at least two kinds of accuracy: what they call 'system accuracy' (a theoretically ideal mental model which may be hard to derive) and 'expert accuracy' (comparisons to an already-trained human expert) (Jonker, Riemsdijk, and Vermeulen 2010).

3.3.3. Metrics accounting for specialization

Specialization is a key aspect of teams, as defined by the field. For fairly complex systems, it is unreasonable to expect any team member to have an accurate mental model of the whole task. Thus, Cooke discusses several additional measures to capture appropriate knowledge for specialized roles. Though these metrics offer compelling theoretical improvements over simple similarity and accuracy measures, they appear to be less well represented in the literature.

Heterogeneous Accuracy tests each team member only on knowledge specific to their role; the team is then evaluated as an aggregate of team members' individual scores.

Inter-positional Accuracy is the opposite of heterogeneous accuracy: it tests team members on the knowledge specific to the roles of the other members. In a study analyzing teamwork from domains as varied as avionics troubleshooting to undergraduate studies, Cooke finds that teams often become less specialized and more inter-positionally accurate with time. Moreover, this increase in inter-positional accuracy is associated with an increase in team performance (Cooke et al. 1998).

Knowledge Distribution is similar to Accuracy but instead measures the degree to which each piece of important information is known by at least one teammate (see below, §3.5.6).

3.3.4. Perceived mutual understanding

Thus far, the metrics discussed are concerned only with the objective characterization of team member knowledge. Other authors, however, emphasize the importance of perception—what team members think about the accuracy and similarity of team knowledge (Burtscher and Oostlander 2019; Rentsch and Mot 2012). Perceived mutual understanding (PMU) 'refers to subjective beliefs about similarities between individuals' as opposed to objective knowledge of their existence, and it has also been correlated with team performance (Burtscher and Oostlander 2019).

3.3.5. Similarity vs. accuracy

There is some inconsistency in the literature regarding which metrics are appropriate for SMMs. Cooke et al. detail multiple types of accuracy measures and highlight their relevance, and several studies use them as their basis; however, although none dispute its relevance, several authors exclude accuracy from their definitions of SMMs. In Jonker et al.'s conceptual analysis, SMMs are defined explicitly in terms of the questions two models can answer and the extent to which their answers agree, eschewing accuracy measures on the basis that they do not necessarily involve comparisons between teammates (Jonker, Riemsdijk, and Vermeulen 2010). DeChurch and Mesmer-Magnus likewise exclude studies that do not explicitly compare teammates' mental models to each other from their meta-analysis of SMM measurement (DeChurch and Mesmer-Magnus 2010).

Intuitively, similarity alone can benefit a team in any situation where an exact protocol is unimportant, so long as teammates can consistently predict each other—for example, the

convention of passing on the right in hallways, or assigning a common name to a certain concept. However, there will always be aspects of a task for which there is an objective, right answer, such as the flight characteristics of a plane or the adherence to laws or procedures. Consequently, there are good reasons one may measure the accuracy of mental models alongside their similarity, whether one considers this an element of the 'shared' mental model or not.

3.4. Development and maintenance of SMMs in human teams

The previous sections discuss what SMMs are and how they can be characterized, but where do they come from? What mechanisms allow them to form? How can we foster their development and prevent them from degrading? Substantially less research addresses these questions (Bierhals et al. 2007), but they are key to our goal of fostering SMMs between humans and machines. In this section, we summarize the relevant literature on the factors and social processes that enable human teams to develop and maintain SMMs.

3.4.1. Development

The nature of SMM development is fundamentally a question of learning. Each human is unique: they approach life with their own mental models of the world. Therefore, for team members' models to become shared, change must occur to establish common ground. This in turn requires interaction with other team members; left in isolation, an individual's models will remain stagnant (Van den Bossche et al. 2011). Mental model development has been studied through various lenses, including verbal and nonverbal communication, trust, motivation, the presence or absence of shared visual representations of the task environment, and a specific set of processes termed 'team learning behaviors.'

Verbal communication is the most obvious form of team member interaction that impacts the development of SMMs. The two are mutually reinforcing: communication can lead to the formation of better SMMs, and the existence of better SMMs can also facilitate better communication (Bierhals et al. 2007). This will result in either a positive or negative feedback loop: a good SMM will result in constructive communication which will result in a better SMM, and vice versa.

Many studies agree that, while verbal communication is important, far more is said nonverbally—as much as 93 percent of all communication, according to Mehrabian's mid-century research (Mehrabian 1972) (though this exact figure has been disputed (Burgoon, Guerrero, and Manusov 2016)). This includes facial expressions, posture, hand gestures, and tone (Burgoon, Manusov, and Guerrero 2021). Bierhals et al. 2007 study of engineering design teams, and Hanna and Richards 2018 study of humans and intelligent virtual agents (IVAs) explore the effects of communication on SMM development (Bierhals et al. 2007; Hanna and Richards 2018). Both verbal and nonverbal aspects of communication are found to significantly influence development of both shared taskwork and teamwork models. Interestingly, nonverbal communication seems to be more important in the development of taskwork SMMs, whereas verbal communication plays a larger role in development of teamwork models (Hanna and Richards 2018).

Beyond communication, higher order social and personal dynamics are at play. These include personal motivation, commitment to task completion, mutual trust, and other emotional processes such as unresolved interpersonal conflicts. Unfortunately, there is a

paucity of research concerning the impact of such factors on SMMs. Bierhals et al. make reference to 'motivational and emotional processes' to account for peculiarities in their results, but their experiments are not designed to specifically measure these factors (Bierhals et al. 2007). An exception is Hanna and Richards' study, which directly correlates the effects of trust and commitment to SMM development in teams of humans and IVAs (Hanna and Richards 2018). (However, these measurements were made only via subjective self-assessments of SMM quality.) Both better shared teamwork and taskwork mental models are found to positively correlate with human trust in their artificial teammate; the effect of sharedness in the taskwork model is found to be slightly stronger than that of the teamwork model. In addition, teammate trust is found to significantly correlate with task commitment, which is found to significantly correlate with improved team performance. From what little research has been done on the topic, it is clear that higher order social and personal dynamics are fundamentally intertwined with SMM development. Further research is warranted to fully discern the nature of this relationship.

Some studies suggest that mutual interaction with shared artifacts, such as shared visual representations of the task environment, can be used to facilitate SMM convergence. Swaab et al.'s study on multiparty negotiation support demonstrates that visualization support facilitates the development of SMMs among negotiating parties (Swaab et al. 2002). Bolstad and Endsley's study on the use of SMMs and shared displays for enhancing team situation awareness also suggests that shared displays help to establish better SMMs, and correlates the presence of such shared artifacts with improved team performance (Bolstad and Endsley 1999). These findings support the common assertion that mental models are frequently pictorial or image-like, rather than script-like in a language processing sense or symbolic in a list-processing sense (Rouse and Morris 1986). By providing an accessible domain conceptualization in the form of a shared visual display, it is intuitive that team mental models would converge around the available scaffolding.

Scaffolding is also insightful with respect to the role of prior knowledge in mental model development. The term 'scaffolding' as a metaphor for the nature of constructive learning patterns originated with Wood, Bruner and Ross (Wood, Bruner, and Ross 1976) in the mid 1970s and has resonated with educators ever since (Hammond 2001). The basic idea is that educators add support for new ideas around a learner's pre-existing knowledge structures. As the learner gains confidence with the new ideas, assistance or 'scaffolding' can be removed and the learning process repeated. An educator's ability to add new information to a learner's knowledge base is largely dependent on the quality of the learners pre-existing schema. Mental models for system control, human-robot interaction, military operations, etc. function in a similar manner. If pre-existing mental models are accurate and robust, they provide a firm starting point for learning new concepts and interpreting new material (Converse, Cannon-Bowers, and Salas 1993; Rouse and Morris 1986). However, if they are inaccurate, containing information gaps and incoherencies, research suggests that they can be difficult to correct (Converse, Cannon-Bowers, and Salas 1993; Rouse and Morris 1986). The process of SMM development is thus highly dependent on the experience, education, socio-cultural background, and moreover, pre-existing mental models of each team member. However, despite the apparent importance of team member prior knowledge and experience level, comprehensive analysis of the effects of these factors with regard to SMM development in human-agent teams has not yet been performed.

Team learning behaviors—broadly categorized as construction, collaborative construction (co-construction), and constructive conflict (Van den Bossche et al. 2011)—are a convenient way of mapping common behaviors seen in teams to various levels of SMM development. They are formally defined in the literature as 'activities carried out by team members through which a team obtains and processes data that allow it to adapt and improve' (Edmondson 1999). Research on team learning behaviors is fundamentally informed by theory on negotiation of common ground, an idea originating in linguistics in 1989 (Clark and Schaefer 1989) and swiftly adapted and expounded on by subsequent learning science researchers (Beers et al. 2007). In theory on negotiation of common ground, communication is viewed as a negotiation, with the ultimate goal of establishing shared meaning or shared belief. Communicators engage in a cyclic process of sharing their mental model, verifying the representations set forth by other team members, clarifying articulated belief statements, accepting or rejecting presented ideas, and explicitly stating the final state of their mental model—whether changed or unchanged (Beers et al. 2007). Construction is the first stage of this process—namely, personal articulation of worldview (Beers et al. 2007).

Co-construction is the portion of the common ground negotiation process primarily concerned with the acceptance, rejection, or modification of ideas set forth by others (Baker 1994). The result of co-construction is that new ideas emerge from the team holistically that were not initially available to each team member (Van den Bossche et al. 2011). If team members accept new ideas and converge around a commonly held belief set, an SMM has been developed. Otherwise, a state of conflict exists.

Constructive conflict occurs when differences in interpretation arise between team members and are resolved by way of clarifications and arguments. Constructive conflict, unlike co-collaboration, was found by Bossche et al. to be significantly correlated with SMM development (Van den Bossche et al. 2011). Conflict shows that team members are engaging seriously with diverging viewpoints and making an active effort to reconcile their representations based on the most current information. Processing of this caliber may just be the very prerequisite for meaningful mental model evolution (Jeong and Chi 2007; Knippenberg, De Dreu, and Homan 2004).

3.4.2. Maintenance

Once an SMM has been established, it will be either maintained, be updated, or begin to degrade. Updating refers to the constant need to keep SMM information current and relevant in light of changing system dynamics, whereas maintenance and degradation refer to the fleeting nature of knowledge in human information processing. All three aspects will here be referred to collectively as *maintenance*. Much like with SMM development, SMM maintenance is largely governed by team member interaction.

In his 2018 PhD dissertation on SMMs, Singh suggests that SMMs are updated through four main processes: perception, communication, inference, and synchronization (Singh 2018). Perception refers to sensing the environment. If team members are in a position to detect the same changes in the environment, SMMs can be updated through perception alone. In most cases, however, communication plays a crucial role in mental model maintenance. Much like with SMM development, team members will communicate to engage in team learning behaviors to co-construct interpretations of evolving system dynamics, and thereby keep their SMM up to date (Van den Bossche et al. 2011). On a slightly deeper level,

inference mechanisms will guide the actual interpretation of perceived system states and verification of communicated beliefs. Though Singh discusses synchronization as a fourth process (Singh 2018), it is debatable whether or not this is fundamentally different from communication. It may be helpful to view synchronization processes as a convenient subclass of communication processes, specifically geared toward maintaining SMM consistency.

Because SMM maintenance is an inherently social activity, higher order social and emotional dynamics such as trust, motivation, and commitment are crucial factors. Trust, for example, plays an essential role in communication. An agent's level of trust in their teammate will directly correlate with the likelihood of adopting that teammate's communicated beliefs (Singh 2018). Research also suggests a correlation between team member motivation for goal completion and SMM maintenance (Wang et al. 2013). This relationship is intuitive; up-to-date SMMs facilitate goal completion: teams that are highly motivated to complete the goal will also be motivated to maintain their SMM. Joint intention theory (Cohen and Levesque 1991) takes this idea one step further. The theory requires team members to be committed not only to goal completion, but also to informing other team members of important events. In other words, team members must be committed to working as a team. A group of individuals could all be highly motivated to achieve a goal, but if they are not also bought into the team concept, an SMM will likely never form—and if it does, it will degrade quickly.

Skill degradation and forgetfulness literature is insightful with respect to the nature of SMM deterioration. Multiple studies have shown that humans retain cognitive skills only for a finite period of time before substantial degradation ensues (Sitterley and Berge 1972; Volz 2018; Erber et al. 1996). This also applies to SMMs of teamwork and taskwork. To maintain shared teamwork models, teams should train together on a regular basis. Research suggests that teamwork cross-training, even when conducted outside of the conventional work domain, can facilitate SMM maintenance (McEwan et al. 2017; Hedges et al. 2019), though some training in the work domain should be retained to maintain shared taskwork models. Maintenance of shared taskwork models can also be aided by regular academic training sessions, as well as pre- or post-task briefings (Dionne et al. 2010; McComb 2007; Johnsen et al. 2017).

3.5. Related concepts

As with mental models, literature on SMMs contains many similar and overlapping terms, and certain terms differ among authors.

3.5.1. Team mental models

Team mental models and SMMs generally refer to the same idea: that effective teams form joint knowledge structures to realize performance gains (Cannon-Bowers and Salas 2001). Team mental models are defined in the early literature as 'team members' shared, organized understanding and mental representation of knowledge about key elements of the team's relevant environment' (Lim and Klein 2006; Mohammed and Dumville 2001; Klimoski and Mohammed 1994). Those same seminal works on Team Mental Models made frequent reference to foundational SMM literature and used the terms interchangeably (Mohammed and Dumville 2001; Klimoski and Mohammed 1994; Converse, Cannon-Bowers, and Salas

1993; Rouse and Morris 1986). Despite their fundamental similarity, subtle differences seem to have emerged in the way the terms are used. The 'sharedness,' for example, of SMMs is indicative of the early belief that similarity among team member mental models was the key determinant of team performance. Later research has demonstrated that additional factors are involved. The use of 'team mental model' de-emphasizes sharedness somewhat, and in turn attempts to place the focus on the study of collective knowledge structures that result from teaming (Cooke et al. 2000). Note that this term is also sometimes used by other authors as a reference to mental models of teams (team models or team interaction models).

3.5.2. Shared cognition

Shared cognition and SMM literature have common origins in the 1990s' and early 2000s' work of researchers such as Cannon-Bowers, Dumville, Mohammed, Klimoski, and Salas (Cooke et al. 2000; Converse, Cannon-Bowers, and Salas 1993; Klimoski and Mohammed 1994; Mohammed and Dumville 2001). Such authors sought to explain the fluidity of expert team decision-making and coordination, hypothesizing that similarity in team member cognitive processing enabled increased efficiency (Cannon-Bowers and Salas 2001). SMM theory is concerned specifically with the theorized knowledge structures underlying this process, whereas shared cognition focuses more broadly on the process itself.

3.5.3. Team knowledge

Team knowledge is a slight expansion upon the idea of team mental models to include the team's model of their immediate context, or situation model (Cooke et al. 2000). The term was introduced and developed primarily by Nancy Cooke. The word choice for 'team knowledge' is specifically selected to specify the study of teams as opposed to dyads or work groups so as to avoid the ambiguity surrounding use of the word 'shared' and to focus study on the knowledge structures at play in teaming instead of broader cognitive processes (Cooke et al. 2000). Team situation models can be seen as more fleeting, situation-specific, dynamic, joint interpretations of the immediate evolving task context; in this way, it is unclear if or how they are distinct from 'team situation awareness', discussed below. Given the concession that SMMs play a primary role in situation interpretation (Cooke et al. 2000; Orasanu 1990), it is not clear how necessary it is to make this distinction. Substantial attention toward situation models is probably only necessary in fast-paced environments such as emergency response, air-to-air combat, or tactical military operations.

3.5.4. Team cognition

Team cognition takes as its premise that cognitive study can be performed holistically on teams and that the study of this emergent cognition is largely a function of team member interaction (Cooke et al. 2013). Though team cognition 'encompasses the organized structures that support team members' ability to acquire, distribute, store, and retrieve critical knowledge,' it is more concerned with the processes than the structures themselves (Fernandez et al. 2017). Thus, SMMs can be viewed as the knowledge structures that facilitate team cognition. Cooke specifies team decision-making, team situation awareness, team

knowledge (including team mental models and team situation models), and team perception as the basic building blocks of team cognition (Cooke et al. 2000).

3.5.5. Distributed cognition

Like shared cognition, distributed cognition suggests that cognitive processes are distributed among team members, but it explores further cognitive distribution in the material task environment—and with respect to time (Hollan, Hutchins, and Kirsh 2000). Distributed cognition stems largely from the work of Edwin Hutchins and has much in common with shared cognition. Both recognize the emergence of higher order cognitive processes in complex socio-technical systems that dwell beyond the boundaries of the individual. Both concepts emerged around the same time and were clearly influenced by some of the same ideas. That said, distributed cognition has its own unique perspective on cognitive processes in complex work teams in socio-technical systems. The task environment in the study of emergent cognitive processes is an important concept with respect to human-automation teaming (Hutchins and Klausen 1996) and as SMM theory is updated to include artificial agents.

3.5.6. Transactive memory

Transactive memory is intimately related to team and distributed cognition and is formally defined as 'the cooperative division of labour for learning, remembering, and communicating relevant team knowledge, where one uses others as memory aids to supplant limited memory' (Ryan and O'Connor 2012; Lewis 2003; Wegner 1987). In other words, not every fact has to be known by every team member, so long as everyone who needs a fact knows how to retrieve it from someone else on the team. Research on this topic conceives of formally analyzing the flow of information between team members as storage and retrieval transactions, analogous to computer processing.

3.5.7. Team situation awareness

Situation awareness may be discussed at the team level, and its relationship to SMMs is analogous to its relationship to mental models at the individual level. As with individual SA, team (or Shared) SA is supported by information availability and adequate mental models. Additionally, team SA is supported by appropriate communication between team members (Entin and Entin 2000).

It is worth noting that many questions used to measure SMMs actually measure SA—that is, rather than asking about general knowledge, experiments may query knowledge inferred about the immediate situation as a proxy for measuring the underlying models. For example, one may ask a pilot and copilot what the next instruction from air traffic control (ATC) will be to test of their knowledge of ATC procedures. This is not an incorrect way to measure SMMs — however, in doing so, the researcher must be aware of other factors that influence SA, such as equipment reliability and the limits of attention and working memory, and consider ways to control for these (Endsley 2004, 13-29).

4. SMMs in human-AI teams

Recently, researchers have suggested that SMMs are a useful lens to study and improve performance in teams of human and AI agents. That is, human and AI teammates can

be led to make accurate predictions of each other and of their shared task and thus achieve better coordination. Attempting to predict human behaviors or intentions and to modify automated system behavior accordingly is an idea that has been investigated previously in the dynamic function allocation (DFA) literature as well as the adaptive automation (AA) literature (Parasuraman et al. 1991; Morrison and Gluckman 1994; Kaber et al. 2001; Rothrock et al. 2002; Scerbo, Freeman, and Mikulka 2003; Kaber and Endsley 2004; Coye de Brunelis and Le Blaye 2008; Kim, Lee, and Johnson 2008; Pritchett, Kim, and Feigh 2014; Feigh, Dorneich, and Hayes 2012b), and in some work on AI expert systems (Mueller et al. 2019, Section 5). In both DFA and AA, the goals were similar, i.e. to adapt automation to better support humans in their work, the mechanisms were slightly different. DFA and AA generally were based on observing human behavior, inferring workload, and adapting accordingly to keep workload at an acceptable level in a very predictable (or at least so it seemed to the system designers) way. In some cases performance was included as part of the objective function, but often workload bounding was the main driver. By contrast, the idea behind SMMs is more general: that in order to function as a team, both humans and automated agents must have an understanding of shared and individual goals, likely methods appropriate to achieve them, and teammate capabilities, information needs, etc. - and that from this understanding, each may seek to anticipate the behavior of the other and adapt appropriately to support joint work.

However, though a handful of studies have approached Human-AI SMMs, no formulation of the concept has yet been supported with a quantitative link to improved team performance. The challenges center around (1) how the AI's mental model is conceived of and implemented, (2) how the Human-AI SMM is elicited and measured, and (3) what factors lead to effective formation of a Human-AI SMM.

In the remainder of this work, we hope to lay the groundwork for this future research. This section reviews existing work on the topic; then, stepping back, we present a broad model of the dynamics of SMM formation in a human-AI team. We then use this to detail the theoretical and practical challenges of applying SMM theory in the context of human-AI teams. In the following sections, we highlight various bodies of literature that may address these challenges—including some that have not yet been discussed in terms of SMMs. Finally, we present general considerations for the construction of experiments to evaluate SMMs in human-AI teams.

4.1. Prior work

In 2010, Jonker et al. presented a conceptual analysis of SMMs, specifically formulated to allow human and AI agents to be considered interchangeably (Jonker, Riemsdijk, and Vermeulen 2010). They emphasized SMMs as being fundamentally defined by (1) their ability to produce mutually compatible expectations in each teammate and (2) how they are measured. They also advocated the strict use of similarity, not accuracy, in metrics of SMMs. That is, if team members answer questions about the task and team the same way, then their mental models are (according to Jonker) shared.

Following this conceptual work, Scheutz et al. proposed (but did not implement) a detailed computational framework in 2017 to equip artificial agents (specifically robots) for SMM formation (Scheutz, DeLoach, and Adams 2017). Gervits et al. then implemented this framework between multiple robots and tested it in a virtual human-robot

collaboration task, claiming in 2020 the first empirical support for the benefits of SMMs in human-agent teams (Gervits et al. 2020). Another study, by Hanna and Richards in 2018, investigated a series of relationships between communication modes among humans and AI, SMM formation, trust, and other factors, and positive correlations between all of them and team performance were found (Hanna and Richards 2018). Zhang proposed RoB-SMM, which focused particularly on role allocation among teams, but only tested the model among teams of artificial agents (Zhang 2008).

Although the experimental results among the above studies are encouraging, they far from fully operationalize the SMM concept. Most critically, both Scheutz and Gervits, one building on the other, discuss SMMs as something possessed by an individual teammate (Scheutz, DeLoach, and Adams 2017, 5), which is inconsistent with their definition in foundational literature (Converse, Cannon-Bowers, and Salas 1993; Jonker, Riemsdijk, and Vermeulen 2010) as a condition arising between teammates when their knowledge overlaps. While Scheutz et al. describe a rich knowledge framework for modeling a human teammate, they make no mention of how this can be compared to the *actual* knowledge state of the human to establish whether this knowledge is truly shared—a shared mental model may or may not exist, but there is no mechanism to test this. Likewise, Gervits et al. make only a qualitative distinction between systems that either did or did not possess an SMM-inspired architecture, rather than a quantitative measurement of SMM quality as described above § 3.3. Their study deals only with SMMs between robot agents, not between robot and human agents; the human's mental model is not elicited, and sharedness between robot MMs and human MM is not claimed. They use the terms 'robot SMM' and 'human-robot SMM' to make this distinction, deferring measurement of the latter to future work. The Hanna and Richards study did produce metrics for SMM quality, but only in the form of introspective Likert ratings about the presence of shared understanding, rather than explicit surveys or behavioral measurements of team member expectations (Hanna and Richards 2018).

4.2. Conceptual model of human-AI SMMs

We present here a conceptual model of the components and relevant interactions in a human-AI team that are thought to lead to the formation of an SMM (Figure 1). Our aim here is to summarize the important concepts, to highlight connections to other bodies of research, and to serve as a point of departure for future debate and refinement of the concept.[2]

In the simplest case, the system involves two team members—one human and one artificial—their shared task environment, and a researcher. We include the researcher in our model because of the significant role they play in determining the SMM using current elicitation methods and to illustrate the challenges and specific considerations of eliciting mental models from both parties during experimentation. The researcher here is an abstraction for any human involved in designing or implementing the human-AI system; outside of experimental settings, the researcher is replaced with a team of system designers or engineers, but the dynamics described here remain applicable. We focus on a dyad team in this example, but these relationships could be extended pairwise to larger groups (see below).

Each teammate's mental model may be loosely divided into its task and team components or viewed holistically. Each individual's model is informed by at least four factors: (1) their

prior knowledge, (2) the task environment, (3) the researcher/system designer, and (4) their teammates. In successful SMM formation, as the teammates perform their task, collaborate with each other, resolve conflicts, and engage in a breadth of team learning behaviors, their mental models converge such that each makes equivalent predictions of the task environment and of each other. The portions of their mental models that produce these equivalent expectations are then defined as the SMM.

Prior experience in a human teammate refers to whatever knowledge, habits, skills, or traits the human brings to the task from their life experience. For artificial agents, this factor may not be applicable in an experimental setting, or it may overlap with the inputs given by the researcher.

Both human and artificial agents may learn from the task environment through direct interaction or explicit training. In an AI specifically, this may take the form of reinforcement learning, or the agent may have a 'fixed' understanding of the task, having been given it *a priori* by its designer as part of a development or training phase.

The researcher(s) play a strong role in both agent's mental models. To the human, the researcher provides training and context that informs how the human will approach their task. They have an even greater influence over the AI (even outside of experimental settings), potentially designing its entire functionality by hand. For any learning components of the AI, the researcher will likely provide 'warm start' models derived from baseline training data, rather than deploying a randomly initialized model to interact with the human.

In an experimental setting, the researcher takes steps to elicit and measure the mental models of each teammate, to assess the quality of their shared mental model or whether they have learned what was intended. This process is nontrivial and requires different methods for human and artificial agents (4.3.2).

Finally, the human and AI acquire mental models of each other through mutual interactions. While some work exists on informing humans' models of AI (e.g. (Bansal et al. 2019)), and some exists on forming AI's models of humans (see §5), little work yet exists on how the social, collaborative processes involved in human SMM formation discussed in §3.4.1 map to the human-AI setting. Furthermore, these interactions are heavily constrained and mediated by the human-machine interface (see §4.3.3). Trust is a major factor in this aspect of the relationship; for more general reviews on trust regulation in human-automation interaction, see French, Duenser, and Heathcote (2018) and Lee and See (2004).

In larger groups, the above relationships are replicated for every pair of teammates which share responsibilities. Between pairs of humans, some elements such as the human-machine interface are omitted, but most factors remain: each human must learn from the task environment, any provided training, and experience working with their teammate. Between pairs of AI agents, opportunities exist for direct communication or for sharing knowledge directly through shared memory, as seen in Scheutz, DeLoach, and Adams (2017). In these cases, the boundaries between agents are blurred and care must be taken in analyses deciding whether to model a team of artificial agents or a single, distributed agent.

4.3. Challenges relative to human-human teams

The literature currently lacks a thorough exploration of what it means to apply the concept of a mental model to an artificial system. By hypothesis, the concept may be applied equally

to AI as to humans: whatever expectations a system produces about its environment or teammates that can lead to useful preemptive behavior is relevant to SMM theory. Nevertheless, all contemporary AI falls into the category of 'narrow' (as opposed to 'general') intelligence; it does not possess the full range of cognitive abilities of a human. All relevant capabilities of the system must be created explicitly. Following from the above model, we now note some additional considerations that AI 'narrowness' necessitates in the application of existing SMM literature to mixed teams of humans and AI, and, where possible, we link to the bodies of literature that may address them.

4.3.1. Forming the AI's mental models

A fully-fledged Human-AI SMM needs an AI with both a task model and a team model. The components of an AI enabling it to perform its core job (task) can roughly be considered its 'task model' - that is, the components of a system that explicitly or implicitly represent knowledge of the external world to be interacted with. Virtually any kind of AI could underlie this portion of the artificial teammate. We consider some of the implications this has for the design and study of a Human-AI SMM in §7.2, but in general the processes for creating and optimizing the task AI are beyond the scope of this review.

Generally speaking, an AI designed for a specific task lacks anything that might be described as a 'team model'; it makes no inferences about the humans using it. However, there have been a number of efforts in computer science and computational psychology to model humans that may be useful to form the team model component of a Human-AI SMM. We detail these in §5.

4.3.2. Eliciting and measuring human-AI SMMs

Although it is generally assumed that humans are able to answer questions about their own knowledge and reasoning, this ability must be specifically designed for in AI systems. In some types of AI, such as the logical knowledge bases used in the work of Scheutz, DeLoach, and Adams (2017), this may be as simple as outputting a state variable. In others, however, particularly deep learning systems, the ability to explain a prediction or engage in any kind of metacognition is a major, largely unsolved problem. This is the subject of the explainable AI field, which we detail further in section 6.

AI's lack of explainability constrains how mental models can be elicited in human-AI teams. Whatever elicitation method is used to measure the mental model, it must be applicable to both a human and an AI. An AI certainly cannot easily complete a Likert survey, nor participate in an interview; observation methods only apply if the AI has something intelligible to observe. Conceptual methods may be appropriate for AIs built on logical databases but not for others. Surveys asking direct, objective questions about the state of the task and team are likely the most accessible for both types of teammate. Process tracing is also an option, depending on how the task environment is set up. Eye tracking, for example, would make no sense for an AI, but interactions with controls might. This problem of elicitation is still open; to date, no study has collected measurements from both the human and AI members of a team.

Another issue is that whereas humans have a very broad ability to adapt their mental models, in any AI system, there is much that is hard-coded and inflexible, both about how the system performs its job and how it interacts with its teammates. For example, there

might be a fixed turn-taking order or frequency of interactions with the system, a fixed vocabulary of concepts, or a limited set of possible commands or interactions. There may be assumptions about the task environment or teammate behavior that are imposed by the AI's engineers, either to reduce the complexity of the implementation or because they are assumed to be roughly optimal. (One can consider these situations as corresponding to fixed elements of the AI' s team or task model, respectively.) This means that in a human-AI team there are many aspects of team behavior for which *similarity* of mental models is no longer sufficient; in all cases where the AI's behavior is fixed, there is a correct set of expectations the human must adopt, and the question becomes more so one of *accuracy*, in which the AI's fixed behaviors are the baseline.

Additionally, including artificial teammates may prompt the modeling and measurement of additional factors that would not ordinarily be considered part of any mental model. For example, Scheutz et al. model cognitive workload of human teammates by monitoring heart rate and use it to anticipate which teammates need additional support (Scheutz, DeLoach, and Adams 2017). It is easy to imagine other human traits that could be modeled, such as personality and tendencies or preferences that the human themselves may not even be aware of. Equivalent traits for machines could include processing bottlenecks and any sensor or equipment limitations not explicitly modeled in the AI itself. These traits fall somewhat outside the scope of SMMs as models that produce *equivalent* expectations, since they exist only in the mental model of one teammate. Nevertheless, they can be useful to model and for fostering performance. (It is a semantic matter for others to debate whether modeling these traits implies a necessary extension of the SMM concept in the human-AI domain or highlights the breadth of factors besides SMMs that impact human-AI teaming.)

4.3.3. The role of the human-machine interface in developing and maintaining SMMs

To develop a successful SMM, team members must be able to acquire knowledge on what their teammates' capabilities and responsibilities are and collaborate with them toward a common goal. In human-human SMMs, this is largely done through multiple modes of communication, as highlighted in section 3.4.1. For human-AI teams, the communication with and acquisition of information regarding other teammates is informed by and largely dependent on the interface/communication module used to display, acquire, and pass information.

A common mode of interaction between a human user and an automated system is through a visual user interface. Many studies have assessed the correlation between the quality of visual user interfaces and user performance, most of which indicate that a poor user interface leads to poor performance (Sutcliffe, Ennis, and Hu 2000; Finley 2013; Osman, Ismail, and Wahab 2009). Several guidelines and best practices have been developed to improve graphical user interfaces (GUIs) to increase usability and retention of use (Stone et al. 2005; Sharp, Rogers, and Preece 2007). The level of abstraction of information has been found to correlate with performance (Sutcliffe and Patel 1996; Janzen and Vicente 1997), where the 'correct' level of abstraction is usually task-dependent or dependent on mental workload limits (Burns, Thompson, and Rodriguez 2002). Task-specific visualizations, such as the timeline metaphor (Plaisant et al. 1996), have proven to be effective in developing and maintaining the individual's mental model of the task at hand. The effectiveness of the design of the interface extends to the development of SMMs, since many use collaborative visualization (Swaab et al. 2002; Siemon et al. 2017) to encourage the convergence of mental models around a shared visual display.

Once an artificial team member is incorporated, the interaction medium must not only be effectively understandable and usable by human team members, but also facilitate meaningful communication and interaction between human and artificial team members. To do this, some studies have investigated multimodal communication approaches between humans and intelligent agents to best mimic the range of communication methods humans have. For example, one approach investigated is to use virtual characters or physical embodiments (robots) to give artificial agents a 'face.' Krämer discusses how virtual characters or avatars are capable of providing human-like verbal (textual or auditory) and nonverbal (facial expressions or gaze) signals (Krämer 2010). Salem et al. showed that robot gestures combined with auditory speech were evaluated more positively than robots who communicated through auditory speech alone (Salem et al. 2012). The modes used to communicate agent beliefs, desires, and intentions (BDI) will play a role in how effectively the human teammate understands the agents' BDI and, therefore, in the quality of the shared mental model. Yusoff and Salim review the effectiveness of collaborative visualization (CoVis) (Isenberg et al. 2012) in developing an SMM and add that 'SMM processing can be increased using visualization and multimedia output capabilities through sophisticated multimodal interaction' (Yusoff and Salim 2020).

Hanna and Richards investigate the impact of multimodal communication (verbal and nonverbal) on the development of an SMM (Hanna and Richards 2018). They showed that an anthropomorphized agent visualized as an avatar was impactful on the team dynamic, and led human users to have human-like expectations from it—for better or worse (Perzanowski et al. 2001; Lyons and Havig 2014). It should be noted that although teammates were working collaboratively, they were working asynchronously and were able to observe other teammates in action to gather information on their behavior/progress/goals. If the team were working synchronously, the communication and SMM maintenance would have to be dynamic rather than procedural. The results showed that verbal and nonverbal communication methods between humans and Intelligent Virtual Agents (IVAs) were significantly positively related to human-IVA teamwork SMMs, indicating that multimodal communication between humans and AI may result in richer, better-quality SMMs.

5. Computational methods toward team models in artificial agents

Various research across computer science and human factors have examined predicting or adapting to the needs of a human collaborator. Though very little of it has used this terminology, all are examples of direct or implicit artificial team models. Since most artificial systems can be considered to have a task model by definition (see 4.3.1), the methods described here address an essential challenge in creating Human-AI SMMs.

Modeling humans is still very much an open problem, and more work exists than can be covered in the current scope. Relevant fields include human-AI teaming (Walsh and Feigh 2021), neuroscience (Das Chakladar and Chakraborty 2018; Chakraborti et al. 2017), human-robot interaction (Gombolay, Wilcox, and Shah 2018), and machine learning (Yokoya et al. 2007). Instead, our aim in this section is to summarize the most relevant research directions to creating Human-AI SMMs and illustrate how they can be discussed in terms of the SMM framework, as well as to highlight areas that have not previously been discussed in these terms.

5.1. Frameworks designed around SMM literature

Some methods for creating artificial SMMs are specifically developed to address the language and key concepts of mental model and SMM literature—specifically, the capacity to describe, explain, and predict the environment—and explicitly model both the task and the team itself (Scheutz, DeLoach, and Adams 2017; Jonker, Riemsdijk, and Vermeulen 2010). Such methods have shown some human-agent team performance benefits in small-scale testing, but much of the work so far is theoretical and the implemented systems, like other historical expert system implementations, have experienced difficulty with robustness and scalability (Mueller et al. 2019, Section 5). Although the mental model and SMM literature does have important insights on the nature of human cognition, other seminal contributions have come from fields such as cognitive science, learning science, and psychology in the study of topics such as schema, scripts, and knowledge representation.

5.2. Dynamic function allocation and artificial team expectation models

Function allocation refers to any principled way of dividing work among team members and is well-studied - see Luo, Chakraborty, and Sycara (2015) for a recent example. Dynamic Function Allocation (DFA) was the idea that team functions could be allocated dynamically (changing over time) so that the human and the automated systems (mostly before intelligent or automated agents were available) either achieved higher joint performance, or that human workload remained within an acceptable range. DFA was often intermixed with the idea of Adaptive Automation (AA) - automated systems that would dynamically adapt usually to improve/maintain performance or contain human workload. DFA was more explicit about changing both the functions given to the humans and to the automation, whereas AA focused mostly on explicit changes to the automation itself; the human adaptation was implicit. Regardless, they often led to similar results as changing the automation necessarily means changing function allocation. Feigh, Dorneich, and Hayes (2012a) summarized the approaches both to triggers of AA as well as the changes to the automation itself, plus critiques of the various approaches to both AA and DFA. Feigh & Pritchett (2014) summarized function allocation approaches.

The results of AA and DFA research were mixed (Parasuraman et al. 1991; Morrison and Gluckman 1994; Rothrock et al. 2002; Scerbo, Freeman, and Mikulka 2003; Kaber and Endsley 2004). In general, it was found that dynamically changing either the way the automation works or the way that the humans were supposed to work with automation was challenging. It is difficult to know when to make the transition in a way that is not disruptive, to know how to make the transition, and which meta-cognitive elements are conducive to it. The broad consensus was that it should be done sparingly and that transitions should be well explained, made clear to the human in real-time, and the human should be trained to understand them. Also, the best results tended to come from automating lower-level tasks (Kaber and Endsley 2004). The shear combinatorics of the various types of triggers and adaptations and their interaction with specific task domains has made the development of a simple model or theory for how to implement DFA or AA challenging. However, AA has made its way into the everyday electronics and human interfaces computer software after significant investment in user testing.

Generally, DFA and AA designs did not include any active model of their human counterparts; rather, scientific results or mental model elicitation was used to inform the design of a fixed algorithm.

5.3. Cognitive load determination

A key aspect of optimal teamwork is determining the moment-to-moment capabilities of a given agent, so that teammates can anticipate each other's needs and effectively distribute work. For robots and virtual agents, these are rather easily deduced (e.g., available processing capability, remaining battery life, or physical characteristics of the robot itself, including available torque, top speed, mass, and dimensions), and physical capabilities of human agents are well documented (e.g., maximum lifting capability, maximum allowable heart rate, top speed). However, determining human cognitive workload is not so straightforward. Though some workload questionnaires are well-accepted, like NASA's TLX, others find that questionnaires are unreliable, difficult to obtain in real time, and often blurred by bias (Fan and Yen 2011). Psycho-physiological techniques for measuring cognitive workload, such as measuring heart rate, eye movements, and brain activity, have shown some success but are confounded by non-workload–related environmental factors (Scheutz, DeLoach, and Adams 2017).

There is some evidence that performance on a trivial secondary task, such as pressing a button periodically or performing memory tasks, can be used as an indicator of cognitive workload: as cognitive load increases, attention must be focused more on primary objectives, and secondary task performance generally degrades. One difficulty of this metric is the challenge of choosing an appropriate secondary task that is difficult enough to provide useful insight while not detracting too much from primary task performance. Despite this, one system employing hidden Markov models for cognitive load determination has shown particularly great success in certain task domains (Fan and Yen 2011). Such work is a crucial step in enabling artificial awareness of a user's cognitive state, and similar methods may be capable of eliciting more information about the user's holistic mental model.

Cognitive load would most accurately be described not as an element of a mental model, but as something inferred by a teammate model. Cues from the human, such as the secondary task performance used by Fan and Yen, are used by the model to infer a cognitive load state. This state can be further used by the teammate model to make predictions about how the human's behavior or needs will change in response to different load levels, such as, in the above case, determining when to exchange information.

5.4. Theory of mind representation

Theory of mind is a concept from psychology that refers to the ability of humans and other animals to infer and deduce the cognitive state of another agent, including their beliefs, intentions, and desires (Premack and Woodruff 1978). The psychological or algorithmic process of theory of mind representation results in a recursive structure whereby the mental model of an agent contains a model of the mental model of their teammate, which contains a model of how the teammate is modeling them, and so on (Bosse, Memon, and Treur 2011; Gmytrasiewicz and Durfeet 1995). Such recursive layering of knowledge structures is an 'essential characteristic' of schemata and mental models (Rumelhart and Ortony 1977).

Numerous computational efforts have sought to recreate this ability in artificial agents (Mueller et al. 2019, Chapter 5, esp. p56). Implementations are generally forms of Knowledge-based AI, using systems of logical rules to draw conclusions from observations of the world and its agents. (Bob put the money in his pocket. Alice can see Bob, therefore Alice knows that Bob has the money.) Older formal attempts at computational theory of mind showed some success in predicting user behavior and successfully performing social manipulation (Bosse, Memon, and Treur 2011). However, these expert system implementations lacked robustness and were not successfully scalable (Scheutz, DeLoach, and Adams 2017). More recent theory of mind implementations employing meta learning show more promise in robustly eliciting user mental models (Rabinowitz et al. 2018).

While computational theory of mind remains difficult, in principle it is perhaps the approach most closely in line with the SMM construct. Indeed, the most fully-featured SMM implementations to date (Scheutz, DeLoach, and Adams 2017; Gervits et al. 2020) are based on theory of mind. However, it should be reiterated for clarity that an estimated cognitive *state* of a teammate is not the same as a mental model. Rather, theory of mind systems contain a mental model of their subjects within the rules they use to *infer* this state.

5.5. Inverse reinforcement learning

Recently, reinforcement learning (RL) has been an active topic in the AI community. The goal within RL is to output an optimal policy for acting in a task environment based on some explicitly-defined reward function, such as learning to hit a ball with a minimum energy expense. This policy can reasonably be interpreted as constituting, or as being key component of, a task mental model.

However, others suggest that 'the reward function, rather than the policy or the value function, is the most succinct, robust, and transferable definition of the task,' (Abbeel and Ng 2004) giving rise to the subfield of inverse reinforcement learning (IRL). IRL seeks to elicit the reward function employed by a human expert in a task. One way to do this is processing user demonstrations to determine the most probable reward function that would prompt observed state-action pairs. That is, given an example set of behavior, what function identifies - and thus defines - the task that is being performed? Seminal algorithms based on user demonstration include Maximum Entropy IRL (MaxEnt) (Ziebart et al. 2008), Bayesian IRL (BIRL) (Ramachandran and Amir 2007), and Adversarial IRL (AIRL) (Fu, Luo, and Levine 2018).

'Shaping' is another method for reward function elicitation in which the artificial agent is free to explore the task environment while the human user gives positive or negative rewards in real time based on agent performance—much like clicker training for household pets. Seminal algorithms for shaping-based IRL include Training an Agent Manually via Evaluative Reinforcement (TAMER) (Knox and Stone 2009) and COrrective Advice Communicated by Humans (COACH) (Celemin and Ruiz-del Solar 2015).

Inverse Reinforcement Learning is generally studied as a means of Learning from Demonstration (LfD); that is, a way to instill AI with expertise in tasks that cannot be easily programmed nor defined by a hand-coded reward function, but which can nevertheless be demonstrated reliably by humans. In this sense it is a way to create task models in artificial agents. Yet, it also points to possibilities for creating teammate and teamwork models: instead of learning a human's policy for the purposes of imitating it, an artificial agent could

simply use that policy as part of a teammate model, enabling it to predict the human's actions and anticipate its needs.

While the ability of LfD algorithms to uncover human policies resembles some form of mental modeling, it lacks capability for handling a key dimension of complex cognitive work - abstraction and decomposition of tasks. Hierarchical learning algorithms attempt to deal with this limitation, at least on the temporal level, by 'decompos[ing] a long-horizon reinforcement learning task into a hierarchy of subproblems or subtasks such that a higher-level policy learns to perform the task by choosing optimal subtasks as the higher-level actions' (Pateria et al. 2021). Hierarchical Inverse Reinforcement Learning (HIRL) has shown success in learning models of how human demonstrators abstract and decompose tasks from demonstration and leveraging these models for accurate behavior prediction. One study dealing with behavior prediction for autonomous vehicles finds an HIRL implementation to more accurately predict human driving behavior than simple Neural Network and Hidden Markov Model baselines (Sun, Zhan, and Tomizuka 2018). For learning based approaches to truly infer or learn a human's mental model they will need some functionality for dealing with abstraction - hierarchical approaches are a step in that direction.

Despite the promise of learning based approaches toward artificial mental modeling of humans, key limitations exist. IRL and HIRL algorithms have been researched and tested predominantly within the formalized problem space of Markov Decision Processes - transferring these problem descriptions to real world environments is a challenge in and of itself. Many of the algorithms from this branch of research also require building a database of demonstrations to learn from - the specific way in which this database of demonstrations is collected could be limited by the nature of a given task environment or the dynamics of the human-AI team under consideration. One key limitation of hierarchical methods in particular is that only temporal abstraction is enabled. While one could conceive of employing similar methods to generate non-temporal concept-map-like models based on human demonstrations, to the best of our knowledge this has not been attempted or demonstrated. The theoretical implications and demonstrated successes of Inverse Reinforcement Learning towards AI's mental modeling of human teammates show promise, but practical limitations and unsolved engineering considerations still impede their full implementation for Human-AI SMMs

5.6. Challenges and limitations

Modeling human cognition is difficult, even in simplified settings. Most systems designed with human cognition in mind (in fields including human factors, ergonomics, and human-robot/AI interaction), make necessarily restrictive assumptions about their subjects, and a great many do not employ explicit 'models' of humans within the system at all. Rather, if model elicitation is done, it is done at design time by experimenters or engineers, and the results are used to guide design principles for a system with static, built-in knowledge of a human. An example of this is the pre-2010s DFA systems discussed above, which use principles of cognitive science to guide systems with hardcoded rules for interacting with a human.

Systems that do create or use explicit models of humans tend to treat humans generically, such as in the logical production rules used in knowledge-based theory of mind systems. Little account is made for individual differences, tending rather to model humans as

procedural, analytical computers limited only by what information is and is not presented to them. Learned models tend to be constructed offline, with data collected in advance from often many different participants.

Models also tend to be rather simplistic, opting often to represent the mind as a finite state machine (such as the HMMs in Fan and Yen (2011)) or with simple linear models. Though these are often sufficient for the designers' purposes, it should be clear that more powerful representations will be necessary in many situations requiring robust teamwork.

Exceptions to these limitations do exist. Inverse reinforcement learning tends to use more expressive deep neural models. In the decision support field, a recent study was able to categorize humans at run time based on which and how many pieces of information they use to make decisions in a particular task (Walsh and Feigh 2021). A recent work in learning from demonstration was able to learn from a series of demonstrators, capturing both a shared task model as well as the individual" style" of each demonstrator (Chen et al. 2020).

Though progress is being made, substantial work remains to enable learning expressive team models that are specialized to individual humans and adaptable to their behavior over the course of teamwork. This work will be essential to enable the full potential benefits of Human-AI SMMs.

6. Explainable AI as a method for SMM formation and elicitation

Forming a mental model of another agent that is sufficient to predict their actions requires some ability to explain or interpret how or why they make their decisions. Difficulty in enabling this level of understanding is endemic of complex systems, but it is particularly relevant in many modern AI systems—especially in ML.

The subfield of explainable artificial intelligence (XAI) has grown rapidly since the mid-2010s, following closely behind the deep learning boom. XAI is motivated by a host of factors, most prominently the desire for ethical and rational accountability in deployed deep learning systems, but the surrounding psychological theory demonstrates that the process of explanation is also closely related to the formation and maintenance of SMMs. Furthermore, explanation of AI can be equated with 'eliciting a mental model' of an AI system, and therefore XAI techniques will be useful, if not essential, in establishing or measuring SMMs in human-AI teams. In this section, we provide a summary of the major research questions of this subfield and their implications for human-AI teaming.

6.1. When and why do we want explanations?

Any time an important decision is made—by a human or a machine —its users want there to be a traceable, verifiable account of how it was reached. Mathematicians show their derivations, courts keep records, accountants have ledgers. Modern AI systems lack this quality, yet they have nevertheless proliferated into increasingly impactful areas of society over the last decade, raising concerns among ethicists, scientists, and policymakers about the fairness and safety of these systems. For example, if a learned model denies someone a loan, how can we check whether it has used all the information available to it rather than simply relying on the applicant's ZIP code?

Common fields of interest to XAI authors include (Scheutz, DeLoach, and Adams 2017):

- Employment and Finance (loans, insurance, etc.)
- Healthcare (such as disease diagnosis and treatment recommendations)
- Criminal Justice (such as assessing recidivism risk)
- Autonomous driving
- Military

In 2016, the European Union increased pressure by including a 'right to explanation' requirement in its General Data Protection Regulation (GDPR) (Lipton 2018), thereby creating a regulatory demand for XAI if AI is to be used in the EU. Additionally, researchers note other potential benefits to an AI that can explain itself, including the ability to better debug and improve AI systems in research settings or to use AI to generate new scientific knowledge (Adadi and Berrada 2018).

Interest in AI-generated explanations is not new. Throughout the 1970s and 80 s, researchers addressed many of the same issues in the domain of rule-driven expert systems. Systems in this period could generate explanations by translating logical traces of their operations into sentences. Later, more sophisticated systems were able to model and correct the user's own knowledge, eventually leading to intelligent tutoring systems for specific domains, before interest in AI at large waned in the late 1980s. (See Mueller et al. 2019, Section 5 for a detailed history.)

XAI has taken on newfound urgency because of the progress made in ML (deep neural networks [DNNs] in particular) whose models pose several distinct challenges for explainability. Adadi and Berrada (2018) provides a good overview: classical AI systems consist largely of rules and principles hand-designed by a human expert, which are applied sequentially during execution, and thus it is easy to decompose a trace of their work into rationalizable steps. By contrast, the only human input to an ML system is training examples; its internal structure is not required to have any human-understandable meaning. Likewise, because ML systems consist of highly interconnected parameters, they are difficult to meaningfully decompose. Although certain models like linear regression and decision trees allow for some human interpretability, DNNs are generally complex and opaque. Further complicating matters, multiple accurate models can result from a single dataset and algorithm. Lastly, Lipton (2018) points out that predictive power in ML (or any statistical model) can result from any correlation, not just ones with a rational cause.

Researchers have considered the need for explanation both in terms of general accountability and the situation-specific needs of users. In the general case, Lipton proposes that 'the demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real world costs in a deployment setting' (Lipton 2018)—constraints such as ethics and legality that cannot (or should not) be quantified and optimized during model development. Doshi-Velez and Kim make a similar argument: often in complex systems the requirement to certify safe or ethical operation cannot be proven explicitly by modeling every conceivable scenario; instead, as long as the system is interpretable, human supervisors can check its reasoning against these criteria (Doshi-Velez and Kim 2017). Other surveys have addressed the specific reasons an operator may want an explanation at a given moment, and a number of taxonomies have been proposed (Adadi and Berrada 2018; Doshi-Velez and Kim 2017; Hoffman et al. 2019; Lipton 2018; Mueller et al. 2019). Hoffman et al., for example, enumerate various 'explanation triggers,' questions a user might have that prompt a request for explanation, including

questions about how it works or how the user is to interact with it, questions about effort required, and questions about consequences of correct and incorrect usage.

Both types of support are needed in human-AI teaming. The human's role in many teams will be to ensure that the AI functions rationally and meets any external criteria. Likewise, all of the above explanation triggers are relevant in helping the human teammate form an accurate mental model of how their partner works and of their strengths and weaknesses.

6.2. What are explanations and how do they work?

To inform machine-to-human explanation, XAI researchers study explanation between humans, drawing on work in philosophy of science, psychology (cognitive, developmental, social, and organizational), education and training, team science, and human factors. To date, there is still no single agreed upon scientific definition of explanation. Mueller et al. provide a survey of the major conclusions about explanation from philosophy, cognitive science, and psychology, which we summarize here (Mueller et al. 2019).

6.2.1. Explanation is a fundamental part of reasoning

Various psychological papers strongly link explanation to a number of cognitive functions, including: learning (in general); counterfactual reasoning (Why didn't something else happen?); contrastive reasoning (How is X different from Y?); abductive reasoning ('Inference to the best explanation'); analogical reasoning; prospective reasoning (Prediction); and hypothesis testing. It is especially closely related with causal reasoning—often an explanation is the same as a highlighting of causal relationships (Mueller et al. 2019, 18).

6.2.2. Explanations both use and shape mental models

A key role of mental models, by their own conception, is to generate explanations, allow anticipation, and support reasoning. In turn, explanations guide the further development of mental models by proposing causal relationships or directing attention. This is true both in explanations from one agent to another, and in what Mueller et al. (2019, 28, 78, 91) call 'self-explanation,' wherein a person uses their own mental models to create and evaluate explanations for phenomena, which they indicate as a useful learning exercise.

6.2.3. Explanations can be global or local

Mueller et al. (2019, 66, 87) use the terms 'global' and 'local' explanations to differentiate discussions of how a system works in general from those of why a specific decision was made.

6.2.4. Explanations are context- and learner-dependent

There is no way to guarantee that an explanation will be useful to the user without considering why they want one and what they already know. Likewise, many different forms of communication can be useful as explanations. According to Mueller et al., 'the property of "being an explanation" is not a property of text, statements, narratives, diagrams, or other forms of material. It is an interaction of: (1) the offered explanation, (2) the learner's knowledge or beliefs, (3) the context or situation and its immediate demands, and (4) the learner's goals or purposes in that context' (Mueller et al. 2019, 86). Therefore, one might cast any

request for explanation as a desire to confirm, refine, or correct a mental model of some phenomenon, and any communication or stimulus that leads the learner to satisfy this desire can be considered an explanation (p.71).

6.2.5. Explanation is a collaborative, iterative process

Though an explanation may consist of a single query and response, equally often it is an iterative activity where mutual understanding is achieved over multiple, possibly bidirectional question-answer pairs, in which each participant must form a mental model of the other's knowledge.

6.2.6. Explanations can take many different forms

Numerous formats and types of explanations have been proposed. Mueller et al. (2019, 84) write that explanations can take the form of visualizations, text, formal/logical expressions, concept maps, graphs/networks, tables, abstractions, timelines, or hierarchies, and can refer to, for example, examples (of misclassifications, counter-examples, outliers, clear cases, close competitors), patterns, features and parameters, strategies and goals, algorithms and proofs, events, narratives, and cause-effect relationships. A few specific common approaches include heatmaps over which regions of an image or text passage were important to an inference (salience maps (Zeiler and Fergus 2014), or attention (Dong et al. 2020)), selections of similar and dissimilar training examples to a particular input, and automated captioning systems that predict and explain the classification of an image.

6.3. Explanations must be 'good enough' and no better

An important consideration for explanation systems is to tell the user no more than they need, to avoid overloading them with information or wasting time. Taking this idea further, with repeated questionings (Why? Why?) it is likely that any formal explanation process must reach a statement that is unexplainable and *must* be accepted as-is (reminiscent of the incompleteness theorems). In that case, one approach to explanation may simply be to satisfy the user. Mueller et al. and others refer to this as 'justification': any argument about why something may be true, such as citing the general accuracy of the system or pointing to a correlation that is plausible without detailing its cause. Providing a simplified justification may be expedient, and Mueller et al. (Mueller et al. 2019, 88) write that users may even prefer this to tedious precision. Ideally, an explanation should capture the most important information while omitting nuances where appropriate for brevity.

However, this strategy carries risks: cognitive biases may lead a user to accept an insufficient explanation or overlook key facts if too much is simplified. Worse, a system whose only goal is to persuade may serve to disguise biases in a system rather than highlight them (Lipton 2018). This trade-off—accuracy and didacticism versus clarity and persuasion—is one of the key challenges of explanation.

6.4. Explanation vs. transparency

A key theoretical distinction that has gained traction in the field is the difference between transparency and post hoc explanations. Lipton (2018) defines a 'transparent' system as one that a user can analyze, distinct from one that can provide 'post hoc'

explanations after a decision has been made. A system can have one property or both: post hoc explanations can be used to provide insight into a model that is on its own inscrutable, and he argues that this is the case for how humans explain their own thinking.

He further proposes three qualities that might render a system 'transparent':

Simulatability referring to a model whose computations a human can reasonably reproduce

Decomposability referring to a model in which 'each input, parameter, and calculation admits an intuitive explanation'

Algorithmic transparency referring to a training system that is well characterized; for example, linear models can be proved to converge to a unique solution.

Lipton (2018) adds that transparency and post hoc explainability are not absolutely distinct. A transparent system nevertheless allows a user to form self-explanations of it, and the right post hoc explanations could themselves lead to transparency. One analogy could be the difference between being given a bicycle to disassemble and watching a video tutorial on bicycle repair. In this case, we might view them as two ends of a spectrum, corresponding to varying levels of structure or abstraction applied to the information given to the user, between *passive* (transparent) explanation and *active* (post hoc) explanation.

Each strategy presents challenges. In theory, a transparent system should allow a user to construct accurate explanations on their own, but at the cost of more effort (Sørmo, Cassens, and Aamodt 2005) and experience with the system. Certain obstacles may prevent making a system fully transparent—for example, a model that is proprietary or contains private data (Chakraborty et al. 2017), but some level of transparency is usually gained by the human when operating the system or working with it. It may also be impossible to make a sufficiently large model human-simulatable (Lipton 2018), or at least accurately so (Javaux 1998). By contrast, relying on post hoc explanations introduces the risk of overselling a faulty explanation or oversimplifying. Indeed, asking how we know a model-provided explanation is accurate may return us to the original problem.

6.5. How XAI systems are evaluated

Another line of theoretical work concerns the development of common standards of evaluating XAI systems, which will in turn be a necessary step in validating the explainability of artificial teammates. Hoffman et al. (2019) stress the need to evaluate systems in practice with end users and discusses six broad ways an XAI system might be judged:

1. Intrinsic merits of the explanations produced, such as clarity or precision
2. How subjectively satisfying the explanations are to users
3. The effect the system has on the user's mental models
4. How the system influences the curiosity of the user
5. How the system influences the trust of the user
6. The task efficacy of the entire human-machine work system

Doshi-Velez and Kim propose three broad evaluation methods that roughly trade between metric accuracy and practicality of conducting experiments (Doshi-Velez and Kim 2017):

1. **Application-grounded evaluation** Evaluate the system in the real world, alongside human experts, in a real or realistic task.
2. **Human-grounded metrics** Evaluate the system on a toy problem with amateur humans.
3. **Functionally-grounded metrics** Design a mathematically defined metric that approximates the intended use case of the system.

Finally, Mueller et al. provide an extensive literature review of evaluation methods in section 8 of their work (Mueller et al. 2019).

6.6. Implications for SMMs in Human-AI teaming

Any artificial system must be explainable to its human partner, either implicitly (via transparency) or explicitly (via active explanations), in order for a shared mental model to emerge, and the activity of explanation plays multiple roles.

Explanation broadly, as a collaborative, iterative process, is equivalent to the establishment of an SMM: both participants strive to achieve a mutually compatible mental model of some phenomenon, be it a task or anything else, and each must establish a mental model of the other participant's knowledge to be successful. The only difference is that explanation lacks the explicit assumption that both participants are part of a cooperative team.

Explanation also has a potential role in the acquisition of and maintenance of shared knowledge (mental models + situation awareness). Though most examinations of explanation focus on revealing and resolving errors in long-term knowledge (mental models), it can also reveal and resolve errors in short-term facts.

Finally, most techniques to elicit the mental models of humans—structured interviews, verbal protocols, concept mapping, etc—all fit the general definition of explanations, implying that the process of eliciting a human's mental model relies on processes of explanation. This means, conversely, that we might consider XAI methods as techniques to elicit the mental model of black-box AI systems, both for purposes of measuring Human-AI SMMs and conveying knowledge to the human about their teammate.

7. Considerations for the design of Human-AI SMM experiments

Much work remains in order to establish empirical results around Human-AI SMMs, and the space of possible experiments is vast, spanning diverse task domains, numerous varieties of AI, and the still-ambiguous task of measuring SMMs. The task specification, the form and content of the SMM, and choice of AI flavor all depend strongly upon each other and determine what hypotheses can be tested. Here we present a breakdown of the various factors affecting SMM development and of considerations for designing an artificial SMM experiment. Many questions below serve as possible independent variables for study, others

suggest factors to be controlled for; all are useful to consider. Our hope is that the questions outlined below provide a starting point to systematically explore the problem space.

7.1. Work definition

The choice of work domain and task influences the entire study by determining what work there is to be done, how it can be subdivided, and what types of AI are applicable. Russell and Norvig give an excellent taxonomy of task environments from an AI perspective in Chapter 2§3.2 of their book (Russell and Norvig 2020), classifying environments as **fully** or **partially observable**, **single** or **multiagent**, **deterministic** or **stochastic**, **episodic** or sequential, **static** or **dynamic**, **discrete** or **continuous**, and whether the mechanics are **known** or **unknown**. It is probably desirable to pick a task that is neither so complex that the human-AI team can show no success, nor so trivial that a human or simple computer program could solve it independently. It is also worth considering whether results in one task domain will generalize to others of interest.

Once the work domain is chosen, who does what? How is labor divided among the team members? This aspect of team design is often called 'function allocation' (Pritchett, Kim, and Feigh 2014; IJtsma et al. 2019). What set of tasks and moment-to-moment information is specific to each team member, and what is to be shared? At what level of task abstraction is this information exchanged?

Taken together, these definitions of the work domain, function allocation, and team structure guide the technical implementation of the system and form a baseline for each of the four components of the desired shared mental model—the task model and team model of each team member.

7.2. Design of the AI

Two components make up the AI's portion of the SMM: its model of the task and its model of the team. Which of these components are under study, and how are they implemented? Is the task model hard-coded (e.g. as a script or knowledge base), or is it learned? Is team interaction a learned model, or made implicit via some fixed sequence of interactions? Are attempts made to model or adapt to the behavior of a specific user? Of human users in general? What types of AI are chosen for the system's various components—are they easy to analyze and explain, or are they black boxes? These questions inform what sources of information are available to experimenters when the time comes to measure the AI's portion of the mental model. They also may influence how effectively the human is able to form a mental model of the AI's behavior.

7.3. Team interaction

The substrate of a shared mental model is the interactions that occur between teammates, essential for information to be shared and team learning to take place. What means of interaction are available to the team? What interactions take place when?

Human teams can communicate verbally, through writing, or nonverbally through gestures, tone, etc., and although many of these modes are possible with machines, they are

considerably harder (Mohammed and Karagozlu 2021; Miller 2004; Bass et al. 2011). On the other hand, machines enable new forms of communication through GUIs, such as charts, maps, animations, or simple persistent displays. Which modes will the system use?

How does the design of the human-machine interface (§4.3.3) influence the way information is exchanged? Is information made passively available on a screen, or is it exchanged in transactional dialogues? For information that is shared between the team members, what interactions keep it synchronized so that shared situation awareness is maintained? Are any involved GUIs designed in a way that makes appropriate information visible to the human at the right times? For text interaction, are there fixed prompts and responses, or is a more fluid language system available? Does the system have the capability to explain itself to the user? Can the AI ask questions?

How often do interactions occur? Which interactions are initiated by the human and which by the AI? Do interactions follow a set protocol, or are they initiated at-will by one or both team members?

Careful consideration of the interactions between team members will inform decisions about what information and tasks should be present in the SMM definition, and it will also help ensure that the right information actually ends up being shared (Mingyue Ma et al. 2018).

7.4. Prior experience, instruction, and SMM development

In an experimental setting, three sources of domain knowledge can affect how a team member's mental model will evolve: past experience, instruction and training from the experimenter, and practice interacting with the team and work domain. The human's prior experience determines what set of mental models they start with and how easily they can adapt to the experimental system. There appears to be a 'goldilocks zone' between a novice who knows too little and an expert who is set in their ways. This can apply both to one's domain knowledge and their experience with AI in general, and it is closely related to the maintenance of appropriate trust levels (Lee and See 2004). Generally, only the human will have appreciable prior knowledge. However, if the AI uses ML to adapt to the task, the human, or both, then any warmstart data used will fill a similar role.

Much of the early work in SMMs stemmed from the study of team training practices (Converse, Cannon-Bowers, and Salas 1993; Mathieu et al. 2000; Stout et al. 2017). A number of experiments in human teaming use shared training or cross training as the basis for their assumption of the existence of an SMM (see section 3.3). Though group training will not be a factor in systems with a single human, instruction provided by experimenters in general provides initial expectations that will influence how the user's mental model develops. It can also help to fill any gaps in domain knowledge a particular participant might have.

It is therefore important to carefully consider what training will be presented to users, keeping in mind any ways it may bias their use of the system, and to validate any assumptions made. It may also be useful to consider how much instruction is given on how to use the system: is the goal to guide users to a specific interaction pattern or to see what use patterns emerge naturally? Does the system need explanation, or should it be self-explanatory? Will there be secondary rounds of instruction after the initial introduction to the system, or not?

Finally, once the user and AI begin interacting, their SMM will form and develop through an iterative learning process. The human's mental models will generate expectations and

change when these expectations are broken. Any adaptive AI components will update in a similar way. This interplay involves experimentation (by the teammates), the mediation of trust, and explanation where possible (Bansal et al. 2019).

Important design considerations for this phase include the length of time the team is given to practice before measurement begins, whether and how the SMM changes over time in long experiments, and how it may degrade or be maintained.

7.5. Measuring outcomes

A test of the validity of a Human-AI SMM will naturally involve a measure of overall task performance appropriate to the chosen task domain. Also important is the decision whether to explicitly evaluate the SMM or to assume it exists implicitly based on the experimental treatment.

Because measuring SMMs directly is still an open problem, there are many possibilities to consider: (1) Which of the mental models should be measured, and how? (2) Will you test beliefs of team members during operation, e.g., in a modified SAGAT? (3) Can this be done without undue interference with the system? (4) Will you give an assessment after performance with a quiz or survey? (5) Are conceptual methods applicable—can a concept map be derived from the AI? (6) Will the measurements be made of each team member individually, or is it possible to apply them to the team holistically?

Finally, once measurements are obtained, how will they be analyzed? Will the mental models be compared to the 'as-designed' baseline for the system (i.e., measured for accuracy), or to each other (i.e. similarity)? If individual members' mental models are obtained, what aggregating metric will be used to judge the overall SMM?

8. Conclusions

Shared mental model theory has been shown to be useful for understanding, predicting, and improving performance in human teams. Teams which are able to effectively synchronize information and predict each other's behavior are better able to cooperate, and it is reasonable to extend this idea to human-AI teams. However, there are a number of challenges: Measuring SMMs in humans is difficult and there is disagreement within the field, both around terms and the best ways to elicit and measure SMMs. Additionally, many promising SMM measurement techniques remain largely untested by experiments. Both these issues are compounded when considering artificial agents and the challenges of representing and measuring their mental models.

While contemporary research applying SMM theory to human-AI teams has produced some AI systems that are effective at modeling human teammates, none has yet measured an SMM between human and artificial agents in an experimental setting (section 4.1)—an essential final step in supporting SMM theory in this context. Three steps are necessary in order to accomplish this.

First, clear, agreed-upon definitions are needed for SMMs in the human-AI context that are consistent with those in the human context. This is necessary to make clear how prior work on human SMMs may or may not transfer to the human-AI domain and for accurately comparing the results of future work. We propose the model in section 4.2 as a point of departure for this purpose. Second, best practices must be established for comparing knowledge between human and artificial agents, both in terms of elicitation and

metrics, so that SMMs can be rigorously measured in human-AI teams. Many options for these practices are outlined in section 7. Lastly, the framework put forth in this paper provides a useful way to integrate contemporary and future advances in human-modeling and explainability in AI with human factors research, ideally providing a common language between the two fields. These technologies are essential to SMM theory as tools for forming and understanding mental models in artificial agents as deep learning and other 'black box' AI becomes prevalent in human-AI teams, and for communicating such models to human teammates.

Addendum

At the end of the review process of this work, the National Academy of Sciences published its report *Human-AI Teaming: State of the Art and Research Needs* (National Academies of Sciences Engineering and Medicine 2022). We address it here for the sake of completeness. It covers diverse issues of human-AI teaming and comes to many similar conclusions on the needs of the field. The work does not focus on Shared Mental Models, though they are discussed; rather, it focuses on Shared Situation Awareness in human-AI teams. Because of the dual nature between SA and mental models, we view these works as complementary and refer readers to this report for further information.

Notes

1. Here we use the term *AI* to represent the full range of possible automation and autonomy, whether enabled by learning agent such as ML, deep learning, artificial intelligence, or a more static reasoning system.
2. A further model emphasizing the information content, rather than the formation, of this relationship can be found in (of Sciences Engineering and Medicine 2022, 28); see our addendum in §8.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Robert "Bowe" Andrews is a Lieutenant in the United States Air Force and student pilot in the Euro-NATO Joint Jet Pilot Training (ENJJPT) program at Sheppard AFB. He received undergraduate degrees from Georgia Tech in 2020 in Russian and Mechanical Engineering, and a Master of Science in Aerospace Engineering in 2021. During graduate school, Bowe conducted research with the Cognitive Engineering Center studying shared mental models in human-AI teams. He is particularly passionate about the optimization of next generation human-machine aerospace systems.

J. Mason Lilly is an M.S. Computer Science student and research assistant in the Cognitive Engineering Center at Georgia Tech. His chief interests are deep learning applications in robotics and human-AI interaction. He earned his B.A. in Computer Science & Applied Music from Transylvania University in 2016. Following graduation, he will be joining Microsoft in Atlanta in a robotics role. In his spare time, Mason plays clarinet and mandolin with his family's band and practices photography.

Divya Srivastava is a fourth-year Mechanical Engineering Ph.D. student in the Cognitive Engineering Center at Georgia Tech. She is researching how shared mental models can be developed

and maintained for increased performance in human-autonomy teams. Her research interests are in HRI and user-centric product design. Previously, Divya has worked for Sandia National Laboratories, Amazon Robotics, the Office of Naval Research, and NASA's Jet Propulsion Laboratory, all in the general field of human-robot interaction or autonomous vehicles. Divya received her M.S. degree in Mechanical Engineering from Georgia Tech. She received her B.S. degree in Mechanical Engineering (Honors) with thesis option and concentration in Aerospace, along with a minor degree in Computer Science from Rutgers University. In her spare time, Divya enjoys cooking, reading fiction novels, and singing off-key to the radio.

Karen M. Feigh is a professor of cognitive engineering at Georgia Institute of Technology. She earned her PhD in industrial and systems engineering from Georgia Institute of Technology and her MPhil in aeronautics from Cranfield University in the United Kingdom. Her current research interests include human-automation interaction and using work domain analysis in systems engineering.

ORCID

J. Mason Lilly http://orcid.org/0000-0002-4573-9322
Divya Srivastava http://orcid.org/0000-0003-1063-5344
Karen M. Feigh http://orcid.org/0000-0002-0281-7634

References

Abbeel, Pieter, and Andrew Y. Ng. 2004. "Apprenticeship Learning via Inverse Reinforcement Learning." In Twenty-First International Conference on Machine learning - ICML '04. ACM Press, Banff, Alberta, Canada, 1. doi:10.1145/1015330.1015430.

Adadi, Amina, and Mohammed Berrada. 2018. "Peeking inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)." *IEEE Access*. 6 (2018): 52138–52160. doi:10.1109/ACCESS.2018.2870052.

Baker, Michael. 1994. "A Model for Negotiation in Teaching-Learning Dialogues." *Journal of Artificial Intelligence in Education* 5 (2): 199–254.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Walter S. Lasecki, Daniel S. Weld, and Eric Horvitz. 2019. "Beyond Accuracy: The Role of Mental Models in Human-AI Team Performance." *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 7 (1): 2–11. https://www.aaai.org/ojs/index.php/HCOMP/article/view/5285.

Bass, Ellen J., Matthew L. Bolton, Karen M. Feigh, Dennis Griffith, Elsa Gunter, and William Mansky John Rushby. 2011. "Toward a Multi-Method Approach to Formalizing Human-Automation Interaction and Human-Human Communications." In *IEEE International Conference of Systems Man and Cybernetics*. IEEE, Anchorage, AK.

Beers, Pieter J., Paul A. Kirschner, Henny P. A. Boshuizen, and Wim H. Gijselaers. 2007. "ICT-Support for Grounding in the Classroom." *Instructional Science* 35 (6): 535–556. doi:10.1007/s11251-007-9018-5.

Bierhals, R., I. Schuster, P. Kohler, and P. Badke-Schaub. 2007. "Shared Mental Models–Linking Team Cognition and Performance." *CoDesign* 3 (1): 75–94. doi:10.1080/15710880601170891.

Bisantz, Ann., and Emilie. Roth. 2007. "Analysis of Cognitive Work." *Reviews of Human Factors and Ergonomics* 3 (1): 1–43. doi:10.1518/155723408X299825.

Bolstad, Cheryl A., and Mica R. Endsley. 1999. "Shared Mental Models and Shared Displays: An Empirical Evaluation of Team Performance." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 43 (3): 213–217. doi:10.1177/154193129904300318.

Bosse, Tibor, Zulfiqar A. Memon, and Jan. Treur. 2011. "A Recursive Bdi Agent Model for Theory of Mind and Its Applications." *Applied Artificial Intelligence* 25 (1): 1–44. doi:10.1080/08839514.2010.529259.

Burgoon, J. K., V. Manusov, and L. K. Guerrero (2021). *Nonverbal Communication* (2nd ed.). New York, NY: Routledge.

Brunélis, Thierry and Le Blaye, Patrick. 2009. Towards a human centered methodology for the Dynamic Allocation of functions. *Revue d'Intelligence Artificielle*. 23: 503–522. doi:10.3166/ria.23.503-522. https://www.researchgate.net/publication/220578475_Towards_a_human_centered_methodology_for_the_Dynamic_Allocation_of_functions

Burns, Catherine M., Laura K. Thompson, and Antonio Rodriguez. 2002. "Mental Workload and the Display of Abstraction Hierarchy Information." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 46 (3): 235–239. doi:10.1177/154193120204600304.

Burtscher, Michael J., and Jeannette Oostlander. 2019. "Perceived Mutual Understanding (PMU): Development and Initial Testing of a German Short Scale for Perceptual Team Cognition." *European Journal of Psychological Assessment* 35 (1): 98–108. doi:10.1027/1015-5759/a000360.

Cannon-Bowers, Janis A., and Eduardo Salas. 2001. "Reflections on Shared Cognition." *Journal of Organizational Behavior* 22 (2): 195–202. doi:10.1002/job.82.

Celemin, C., and J. Ruiz-del Solar. 2015. "COACH: Learning Continuous Actions from COrrective Advice Communicated by Humans." In 2015 International Conference on Advanced Robotics (ICAR), 581–586. doi:10.1109/ICAR.2015.7251514.

Chakraborti, Tathagata, Sarath Sreedharan, Anagha Kulkarni, and S. Kambhampati. 2017. "Alternative Modes of Interaction in Proximal Human-in-the-Loop Operation of Robots." ArXiv Abs/1703.08930 (2017).

Chakraborty, Supriyo, Richard Tomsett, Ramya Raghavendra, Daniel Harborne, Moustafa Alzantot, Federico Cerutti, Mani Srivastava, et al. 2017. "Interpretability of Deep Learning Models: A Survey of Results." In 2017 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, San Francisco, CA, 1–6. doi:10.1109/UIC-ATC.2017.8397411.

Chen, Letian, Rohan Paleja, Muyleng Ghuy, and Matthew Gombolay. 2020. *Joint Goal and Strategy Inference across Heterogeneous Demonstrators via Reward Network Distillation*, 659–668. New York, NY: Association for Computing Machinery. doi:10.1145/3319502.3374791.

Clark, Herbert H., and Edward F. Schaefer. 1989. "Contributing to Discourse." *Cognitive Science* 13 (2): 259–294. doi:10.1207/s15516709cog1302_7.

Cohen, Philip R., and Hector J. Levesque. 1991. "Teamwork." *Noûs* 25 (4): 487–512. doi:10.2307/2216075.

Converse, Sharolyn, J. A. Cannon-Bowers, and E. Salas. 1993. "Shared Mental Models in Expert Team Decision Making." *Individual and Group Decision Making: Current Issues* 221 (1993): 221–246.

Cooke, N. J., R. Stout, K. Rivera, and E. Salas. 1998. "Exploring Measures of Team Knowledge." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 42(3): 215–219. doi:10.1177/154193129804200307

Cooke, Nancy J., Eduardo Salas, Janis A. Cannon-Bowers, and Renée J. Stout. 2000. "Measuring Team Knowledge." *Human Factors* 42 (1): 151–173. doi:10.1518/001872000779656561.

Cooke, Nancy J., Eduardo Salas, Preston A. Kiekel, and Brian Bell. 2004. "Advances in Measuring Team Cognition." In *Team Cognition: Understanding the Factors That Drive Process and Performance*, 83–106. Washington, DC: American Psychological Association. doi:10.1037/10690-005.

Cooke, Nancy J., Jamie C. Gorman, Christopher W. Myers, and Jasmine L. Duran. 2013. "Interactive Team Cognition - Cooke – 2013". *Cognitive Science* 37 (2): 255–285. https://doi.org/10.1111/cogs.12009

Craik, Kenneth James Williams. 1967. *The Nature of Explanation*. Cambridge. Cambridge University Press. https://philpapers.org/rec/CRATNO

Das Chakladar, Debashis, and Sanjay Chakraborty. 2018. "EEG Based Emotion Classification Using "Correlation Based Subset Selection." *Biologically Inspired Cognitive Architectures* 24 (2018): 98–106. doi:10.1016/j.bica.2018.04.012.

DeChurch, Leslie A., and Jessica R. Mesmer-Magnus. 2010. "Measuring Shared Team Mental Models: A Meta-Analysis." *Group Dynamics: Theory, Research, and Practice* 14 (1): 1–14. doi:10.1037/a0017455.

Dionne, Shelley D., Hiroki Sayama, Chanyu Hao, and Benjamin James Bush. 2010. "The Role of Leadership in Shared Mental Model Convergence and Team Performance Improvement: An Agent-Based Computational Model." *The Leadership Quarterly* 21 (6): 1035–1049. doi:10.1016/j.leaqua.2010.10.007.

Dong, Zhihang, Tongshuang Wu, Sicheng Song, and Mingrui Zhang. 2020. "Interactive Attention Model Explorer for Natural Language Processing Tasks with Unbalanced Data Sizes." In 2020 IEEE Pacific Visualization Symposium (PacificVis), 46–50. doi:10.1109/PacificVis48177.2020.1031.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." arXiv:1702.08608 [cs, Stat]. http://arxiv.org/abs/1702.08608. arXiv: 1702.08608.

Edmondson, Amy. 1999. "Psychological Safety and Learning Behavior in Work Teams." *Administrative Science Quarterly* 44 (2): 350–383. doi:10.2307/2666999.

Einhorn, Hillel, and Robin Hogarth. 1986. "Judging Probable Cause." *Psychological Bulletin* 99 (1): 3–19. doi:10.1037/0033-2909.99.1.3.

Endsley, Mica R. 1995. "Measurement of Situation Awareness in Dynamic Systems." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 37 (1): 65–84. doi:10.1518/001872095779049499.

Endsley, M. R. 2004. *Designing for Situation Awareness: An Approach to User-Centered Design*, Second Edition (2nd ed.). CRC Press. https://doi.org/10.1201/b11371

Endsley, Mica R. 2016. "Toward a Theory of Situation Awareness in Dynamic Systems." In *Human Factors*. Los Angeles, CA: Sage CA. doi:10.1518/001872095779049543.

Entin, Eileen B., and Elliot E. Entin. 2000. "Assessing Team Situation Awareness in Simulated Military Missions." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 44 (1): 73–76. doi:10.1177/154193120004400120.

Erber, Joan T., Irene G. Prager, Marie Williams, and Marisa A. Caiola. 1996. "Age and Forgetfulness: Confidence in Ability and Attribution for Memory Failures." *Psychology and Aging* 11 (2): 310–315. doi:10.1037/0882-7974.11.2.310.

Fan, X., and J. Yen. 2011. "Modeling Cognitive Loads for Evolving Shared Mental Models in Human-Agent Collaboration." *IEEE Transactions on Systems, Man, and Cybernetics. Part B, Cybernetics: A Publication of the IEEE Systems, Man, and Cybernetics Society* 41 (2): 354–367. doi:10.1109/TSMCB.2010.2053705.

Feigh, Karen M., Michael C. Dorneich, and Caroline C. Hayes. 2012a. "Toward a Characterization of Adaptive Systems: A Framework for Researchers and System Designers." *Human Factors* 54 (6): 1008–1024. doi:10.1177/0018720812443983.

Feigh, Karen M., Michael C. Dorneich, and Caroline C. Hayes. 2012b. "Toward a Characterization of Adaptive Systems: A Framework for Researchers and System Designers." *Human Factors* 54 (6): 1008–1024. doi:10.1177/0018720812443983.

Fernandez, Rosemarie, Sachita Shah, Elizabeth D. Rosenman, Steve W. J. Kozlowski, Sarah Henrickson Parker, and James A. Grand. 2017. "Developing Team Cognition: A Role for Simulation." *Simulation in Healthcare: Journal of the Society for Simulation in Healthcare* 12 (2): 96–103. doi:10.1097/SIH.0000000000000200.

Finley, Patrick M. 2013. "A study comparing table-based and list-based smartphone interface usability". Graduate Theses and Dissertations. 13295. https://lib.dr.iastate.edu/etd/13295, https://core.ac.uk/download/pdf/38926044.pdf

Forrester, Jay W. 1971. "Counterintuitive Behavior of Social Systems." *Theory and Decision* 2 (2): 109–140. doi:10.1007/BF00148991.

French, B., A. Duenser, and A. Heathcote. 2018. Trust in Automation - A Literature Review. CSIRO Report EP184082. CSIRO, Australia. https://www.scribd.com/document/532432101/trust-in-automation-a-literature-review-report, https://www.semanticscholar.org/paper/Trust-in-Automation-A-Literature-Review-French-Duenser/92f07d3d1356307decb6e97382ad884d0f62668d

Fu, Justin, Katie Luo, and Sergey Levine. 2018. "Learning Robust Rewards with Adversarial Inverse Reinforcement Learning." arXiv:1710.11248 [cs] (Aug. 2018). http://arxiv.org/abs/1710.11248. arXiv: 1710.11248.

Felix Gervits, Dean Thurston, Ravenna Thielstrom, Terry Fong, Quinn Pham, and Matthias Scheutz. 2020. Toward Genuine Robot Teammates: Improving Human-Robot Team Performance Using Robot Shared Mental Models. In *Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS '20)*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, 429–437.

Gilson, Richard D., Daniel J. Garland, and Jefferson M. Koonce. (Eds). 1994. *Situational Awareness in Complex Systems*. Proceedings to the Center for Applied Human Factors in Aviation (CAHFA) Conference on Situational Awareness in Complex Systems, Orlando, FL, February 1-3, 1993. Daytona Beach, FL: Embry-Riddle Aeronautical University Press. https://apps.dtic.mil/sti/citations/ADA281448

Gmytrasiewicz, Piotr J., and Edmund H. Durfeet. 1995. *A Rigorous, Operational Formalization of Recursive*, Proceedings of the First International Conference on Multiagent Systems, June 12-14, 1995, San Francisco, California, USA. Cambridge, MA: The MIT Press. ISBN 0-262-62102-9, 125–132.

Gombolay, Matthew C., Ronald J. Wilcox, and Julie A. Shah. 2018. "Fast Scheduling of Robot Teams Performing Tasks with Temporospatial Constraints." *IEEE Transactions on Robotics* 34 (1): 220–239. doi:10.1109/TRO.2018.2795034.

Hammond, Jennifer. 2001. *Scaffolding: teaching and Learning in Language and Literacy Education*. Newtown, NSW: PETA, Primary English Teaching Association. OCLC: 1045428892.

Hanna, Nader, and Deborah Richards. 2018. "The Impact of Multimodal Communication on a Shared Mental Model, Trust, and Commitment in Human–Intelligent Virtual Agent Teams." *Multimodal Technologies and Interaction* 2 (3): 48. doi:10.3390/mti2030048.

Hedges, Ashley R., Heather J. Johnson, Lawrence R. Kobulinsky, Jamie L. Estock, David. Eibling, and Amy L. Seybert. 2019. "Effects of Cross-Training on Medical Teams' Teamwork and Collaboration: Use of Simulation." *Pharmacy* 7 (1): 13. doi:10.3390/pharmacy7010013.

Hoffman, Robert R., Shane T. Mueller, Gary Klein, and Jordan Litman. 2019. "Metrics for Explainable AI: Challenges and Prospects." arXiv:1812.04608 [cs] (Feb. 2019). http://arxiv.org/abs/1812.04608. arXiv: 1812.04608.

Hollan, James, Edwin Hutchins, and David Kirsh. 2000. "Distributed Cognition: Toward a New Foundation for Human-Computer Interaction Research." *ACM Transactions on Computer-Human Interaction* 7 (2): 174–196. doi:10.1145/353485.353487.

Hutchins, Edwin, and Tove Klausen. 1996. "Distributed Cognition in an Airline Cockpit." In *Cognition and Communication at Work*. 1st ed., edited by Yrjo Engeström and David Middleton, 15–34. Cambridge: Cambridge University Press. doi:10.1017/CBO9781139174077.002.

IJtsma, Martijn, Lanssie M. Ma, Amy R. Pritchett, and Karen M. Feigh. 2019. "Computational Methodology for the Allocation of Work and Interaction in Human-Robot Teams." *Journal of Cognitive Engineering and Decision Making* 13 (4): 221–241. doi:10.1177/1555343419869484.

Isenberg, Petra, Danyel Fisher, Sharoda Paul, Meredith Morris, Kori Inkpen, and Mary Czerwinski. 2012. "Co-Located Collaborative Visual Analytics around a Tabletop Display." *IEEE Transactions on Visualization and Computer Graphics* 18 (5): 689–702. doi:10.1109/TVCG.2011.287.

Jagacinski, Richard J., and Richard A. Miller. 1978. "Describing the Human Operator's Internal Model of a Dynamic System." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 20 (4): 425–433. doi:10.1177/001872087802000406.

Janzen, Michael E., and Kim J. Vicente. 1997. "Attention Allocation within the Abstraction Hierarchy." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 41 (1): 274–278. doi:10.1177/107118139704100162.

Javaux, Denis. 1998. "An Algorithmic Method for Predicting Pilot-Mode Interaction Difficulties." In 17th DASC. AIAA/IEEE/SAE. Digital Avionics Systems Conference. Proceedings (Cat. No. 98CH36267), Vol. 1. IEEE, E21–1.

Jeong, Heisawn, and Michelene T. H. Chi. 2007. "Knowledge Convergence and Collaborative Learning." *Instructional Science* 35 (4): 287–315. doi:10.1007/s11251-006-9008-z.

Johnsen, Bjørn Helge, Heidi Kristina Westli, Roar Espevik, Torben Wisborg, and Guttorm Brattebø. 2017. "High-Performing Trauma Teams: frequency of Behavioral Markers of a Shared Mental

Model Displayed by Team Leaders and Quality of Medical Performance." *Scandinavian Journal of Trauma, Resuscitation and Emergency Medicine* 25 (1): 109. doi:10.1186/s13049-017-0452-3.

Johnson-Laird, P. N. 1980. "Mental Models in Cognitive Science." *Cognitive Science* 4 (1): 71–115. doi:10.1016/S0364-0213(81)80005-5.

Philip N. Johnson-Laird. 1983. *Mental models: towards a cognitive science of language, inference, and consciousness*. Cambridge, MA: Harvard University Press, 528 p.

Jonker, Catholijn, M. Riemsdijk, and Bas Vermeulen. 2010. Shared Mental Models: A Conceptual Analysis. https://www.researchgate.net/publication/221456658/_Shared/_Mental/_Models/_-/_A/_Conceptual/_Analysis

Kaber, David B., and Mica R. Endsley. 2004. "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. doi:10.1080/1463922021000054335.

Kaber, David B., Jennifer M. Riley, Kheng-Wooi Tan, and Mica R. Endsley. 2001. "On the Design of Adaptive Automation for Complex Systems." *International Journal of Cognitive Ergonomics* 5 (1): 37–57. doi:10.1207/S15327566IJCE0501_3.

Kim, S. Y., S. M. Lee, and E. N. Johnson. 2008. "Analysis of Dynamic Function Allocation between Human Operators and Automation Systems." In in Proceedings of AIAA Modeling and Simulation Technologies Conference and Exhibit. AIAA Modeling and Simulation Technologies Conference and Exhibit, 18–21 August 2008, Honolulu, Hawaii. doi:10.2514/6.2008-6673.

Klimoski, Richard, and Susan Mohammed. 1994. "Team Mental Model: Construct or Metaphor?" *Journal of Management* 20 (2): 403–437. doi:10.1016/0149-2063(94)90021-3.

Knippenberg, Daan van, Carsten K. W. De Dreu, and Astrid C. Homan. 2004. "Work Group Diversity and Group Performance: An Integrative Model and Research Agenda." *The Journal of Applied Psychology* 89 (6): 1008–1022. doi:10.1037/0021-9010.89.6.1008.

Knox, W. Bradley, and Peter Stone. 2009. "Interactively Shaping Agents via Human Reinforcement: The TAMER Framework." In Proceedings of the Fifth International Conference on Knowledge Capture (K-CAP '09). Association for Computing Machinery, New York, NY, USA, 9–16. doi:10.1145/1597735.1597738.

Krämer, Nicole. 2010. "Psychological Research on Embodied Conversational Agents: The Case of Pedagogical Agents." *Journal of Media Psychology* 22 (2): 47–51. doi:10.1027/1864-1105/a000007.

Lee, John D., and Katrina A. See. 2004. "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* 46 (1): 50–80. arXiv:. PMID: 15151155. doi:10.1518/hfes.46.1.50_30392.

Lewis, Kyle. 2003. "Measuring Transactive Memory Systems in the Field: Scale Development and Validation." *The Journal of Applied Psychology* 88 (4): 587–604. doi:10.1037/0021-9010.88.4.587.

Lim, Beng-Chong, and Katherine J. Klein. 2006. "Team Mental Models and Team Performance: A Field Study of the Effects of Team Mental Model Similarity and Accuracy." *Journal of Organizational Behavior* 27 (4): 403–418. doi:10.1002/job.387.

Lipton C. Zachary. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (May-June 2018), 31–57. doi:10.1145/3236386.3241340

Luo, L., N. Chakraborty, and K. Sycara. 2015. "Distributed Algorithms for Multirobot Task Assignment with Task Deadline Constraints." *IEEE Transactions on Automation Science and Engineering* 12 (3): 876–888. doi:10.1109/TASE.2015.2438032.

Lyons, Joseph B., and Paul R. Havig. (2014). Transparency in a Human-Machine Context: Approaches for Fostering Shared Awareness/Intent. In: *Virtual, Augmented and Mixed Reality. Designing and Developing Virtual and Augmented Environments, edited by R. Shumaker and S. Lackey,* VAMR 2014. Lecture Notes in Computer Science, vol 8525. Springer, Cham. doi:10.1007/978-3-319-07458-0_18

Marks, M. A., S. J. Zaccaro, and J. E. Mathieu. 2000. "Performance Implications of Leader Briefings and Team-Interaction Training for Team Adaptation to Novel Environments." *The Journal of Applied Psychology* 85 (6): 971–986. doi:10.1037/0021-9010.85.6.971.

Mathieu, John E., Tonia S. Heffner, Gerald F. Goodwin, Eduardo Salas, and Janis A. Cannon-Bowers. 2000. "The Influence of Shared Mental Models on Team Process and Performance." *The Journal of Applied Psychology* 85 (2): 273–283. doi:10.1037/0021-9010.85.2.273.

McComb, Sara A. 2007. "Mental Model Convergence: The Shift from Being an Individual to Being a Team Member." In *Multi-Level Issues in Organizations and Time*, edited by Fred Dansereau and Francis J. Yammarino. Research in Multi-Level Issues, Vol. 6. Bingley: Emerald Group Publishing Limited, 95–147. doi:10.1016/S1475-9144(07)06005-5.

McEwan, Desmond, Geralyn R. Ruissen, Mark A. Eys, Bruno D. Zumbo, and Mark R. Beauchamp. 2017. "The Effectiveness of Teamwork Training on Teamwork Behaviors and Team Performance: A Systematic Review and Meta-Analysis of Controlled Interventions." *PloS One* 12 (1): e0169604. doi:10.1371/journal.pone.0169604.

McVee, Mary B., Kailonnie Dunsmore, and James R. Gavelek. 2005. "Schema Theory Revisited." *Review of Educational Research* 75 (4): 531–566. https://www.jstor.org/stable/3516106. doi:10.3102/00346543075004531.

Mehrabian, Albert. 1972. *Nonverbal Communication*. New York, NY: Transaction Publishers. https://doi.org/10.4324/9781351308724

Miller, C. 2004. "Human-Computer Etiquette: Managing Expections with Intentional Agents." *Communications of the ACM* 47 (4): 31.

Mingyue Ma, Lanssie, Terrence Fong, Mark J. Micire, Yun Kyung Kim, and Karen Feigh. 2018. "Human-Robot Teaming: Concepts and Components for Design." In *Field and Service Robotics (Springer Proceedings in Advanced Robotics)*, edited by Marco Hutter and Roland Siegwart, 649–663. Cham: Springer International Publishing. doi:10.1007/978-3-319-67361-5_42.

Mohammed, Susan, Richard Klimoski, and Joan R. Rentsch. 2000. "The Measurement of Team Mental Models: We Have No Shared Schema." *Organizational Research Methods* 3 (2): 123–165. . Publisher: SAGE Publications Inc. doi:10.1177/109442810032001.

Mohammed, Susan, and Brad C. Dumville. 2001. "Team Mental Models in a Team Knowledge Framework: expanding Theory and Measurement across Disciplinary Boundaries." *Journal of Organizational Behavior* 22 (2): 89–106. doi:10.1002/job.86.

Mohammed, Yakubu Bala, and Damla Karagozlu. 2021. "A Review of Human-Computer Interaction Design Approaches towards Information Systems Development." *Brain. Broad Research in Artificial Intelligence and Neuroscience* 12 (1): 229–250. doi:10.18662/brain/12.1/180.

Morrison, Jeffery G., and Jonathan P. Gluckman. 1994. "Definitions and Prospective Guidelines for the Application of Adaptive Automation." In Automation Technology and Human Performance Conference.

Mueller, Shane T., Robert R. Hoffman, William Clancey, Abigail Emrey, and Gary Klein. 2019. "Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI." arXiv:1902.01876 [cs]. http://arxiv.org/abs/1902.01876. arXiv: 1902.01876.

Naikar, Neelam. 2005. "A Methodology for Work Domain Analysis, the First Phase of Cognitive Work Analysis." *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 49 (3): 312–316. doi:10.1177/154193120504900321.

National Academies of Sciences Engineering and Medicine. 2022. *Human-AI Teaming: State-of-the-Art and Research Needs*. Washington, DC: The National Academies Press. doi:10.17226/26355.

Norman, Donald A. 1987. *Some observations on mental models. Human-computer interaction: Amultidisciplinary approach*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 241–244. https://dl.acm.org/doi/10.5555/58076.58097

Orasanu, Judith. 1990. *Shared Mental Models and Crew Decision Making*. Princeton, NJ: Princeton University, Cognitive Science Laboratory.

Osman, Aznoora, Mohammad Hafiz Ismail, and Nadia Abdul Wahab. 2009. "Combinging Fisheye with List: Evaluating the Learnability and User Satisfaction," *International Conference on Computer Technology and Development,* 2009, pp. 49–52. doi:10.1109/ICCTD.2009.247.

Parasuraman, Raja, Brian Hilburn, Robert Mol, and Indramani Singh. 1991. "Adaptive Automation and Human Performance: III. Effects of Practice on the Benefits and Costs of Automation Shifts." Technical Report. Naval Air Warfare Center.

Pateria, Shubham, Budhitama Subagdja, Ah-hwee Tan, and Chai Quek. 2021. "Hierarchical Reinforcement Learning: A Comprehensive Survey." *ACM Computing Surveys* 54 (5): 1–35. doi:10.1145/3453160.

Patrick, John, and Nic James. 2004. "Process Tracing of Complex Cognitive Work Tasks." *Journal of Occupational and Organizational Psychology* 77 (2): 259–280. doi:10.1348/096317904774202171.

Perzanowski, Dennis, Alan Schultz, W. Adams, E. Marsh, and Magdalena Bugajska. 2001. "Building a Multimodal Human-Robot Interface." *IEEE Intelligent Systems* 16 (1): 16–21. doi:10.1109/MIS.2001.1183338.

Plaisant, Catherine, Brett Milash, Anne Rose, Seth Widoff, and Ben Shneiderman. 1996. *Lifelines: Visualizing Personal Histories*, 221–227. New York, NY: ACM Press.

Premack, David, and Guy Woodruff. 1978. "Does the Chimpanzee Have a Theory of Mind?" *Behavioral and Brain Sciences* 1 (4): 515–526. doi:10.1017/S0140525X00076512.

Pritchett, Amy R., So Young Kim, and Karen M. Feigh. 2014. "Measuring Human-Automation Function Allocation." *Journal of Cognitive Engineering and Decision Making* 8 (1): 52–77. doi:10.1177/1555343413490166.

Rabinowitz, Neil C., Frank Perbet, H. Francis Song, Chiyuan Zhang, S. M. Ali Eslami, and Matthew Botvinick. 2018. "Machine Theory of Mind." arXiv:1802.07740 [cs] (March 2018). http://arxiv.org/abs/1802.07740. arXiv: 1802.07740.

Ramachandran, Deepak, and Eyal Amir. 2007. "Bayesian Inverse Reinforcement Learning." In Proceedings of the 20th International Joint Conference on Artifical Intelligence (IJCAI'07). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2586–2591.

Rasmussen, Jens. 1979. *On the Structure of Knowledge - A Morphology of Metal Models in a Man-Machine System Context*. Risø National Laboratory. https://orbit.dtu.dk/en/publications/on-the-structure-of-knowledge-a-morphology-of-metal-models-in-a-m

Rentsch, Joan R, and Ioana R. Mot. 2012. "Elaborating Cognition in Teams: Cognitive Similarity Configurations." In *Theories of Team Cognition: Cross-Disciplinary Perspectives*, 145–170. New York, NY: Routledge/Taylor & Francis Group.

Resick, Christian J., Marcus W. Dickson, Jacqueline K. Mitchelson, Leslie K. Allison, and Malissa A. Clark. 2010. "Team Composition, Cognition, and Effectiveness: Examining Mental Model Similarity and Accuracy." *Group Dynamics: Theory, Research, and Practice* 14 (2): 174–191. doi:10.1037/a0018444.

Rothrock, Ling, Richard Koubek, Frederic Fuchs, Michael Haas, and Gavriel Salvendy. 2002. "Review and Reappraisal of Adaptive Interfaces: Toward Biologically Inspired Paradigms." *Theoretical Issues in Ergonomics Science* 3 (1): 47–84. doi:10.1080/14639220110110342.

Rouse, William B. 1977. "A Theory of Human Decision Making in Stochastic Estimation Tasks." *IEEE Transactions on Systems, Man, and Cybernetics* 7 (4): 274–283. doi:10.1109/TSMC.1977.4309702.

Rouse, William B., and Nancy M. Morris. 1986. "On Looking into the Black Box: Prospects and Limits in the Search for Mental Models." *Psychological Bulletin* 100 (3): 349–363. doi:10.1037/0033-2909.100.3.349.

Rumelhart, D., and Andrew Ortony. 1977. "The Representation of Knowledge in Memory." In *Schooling and the Acquisition of Knowledge,* edited by R. C. Anderson, R. J. Spiro, and W. E. Montague, (1st ed.). Routledge. doi:10.4324/9781315271644.

Russell, Stuart J., and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*. 4th ed. Hoboken, NJ: Pearson. /content/one-dot-com/one-dot-com/us/en/higher-education/program.html

Ryan, Sharon, and Rory V. O'Connor. 2012. Social interaction, team tacit knowledge and transactive memory: empirical support for the agile approach. https://ulir.ul.ie/handle/10344/2698 Accepted: 2012-11-30T11:41:59Z.

Salem, Maha, Stefan Kopp, Ipke Wachsmuth, Katharina Rohlfing, and Frank Joublin. 2012. "Generation and Evaluation of Communicative Robot Gesture." *International Journal of Social Robotics* 4 (2): 201–217. doi:10.1007/s12369-011-0124-9.

Sarter, Nadine B., and David D. Woods. 1991. "Situation Awareness: A Critical but Ill-Defined Phenomenon." *The International Journal of Aviation Psychology* 1 (1): 45–57. doi:10.1207/s15327108ijap0101_4.

Scerbo, Mark W., Frederick G. Freeman, and Peter J. Mikulka. 2003. "A Brain-Based System for Adaptive Automation." *Theoretical Issues in Ergonomics Science* 4 (1-2): 200–219. doi:10.1080/1463922021000020891.

Schank, Roger C, and Robert P. Abelson. 1977. *Scripts, Plans, Goals and Understanding: An Inquiry into Human Knowledge Structures*, 248. Oxford, England: Lawrence Erlbaum.

Scheutz, Matthias, Scott A. DeLoach, and Julie A. Adams. 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. doi:10.1177/1555343416682891.

Sharp, H., Y. Rogers, and J. Preece. 2007. *Interaction Design: Beyond Human-Computer Interaction*. United Kingdom: Wiley. https://books.google.com/books?id=kcEZAQAAIAAJ.

Siemon, Dominik; Redlich, Beke; Lattemann, Christoph; and Robra-Bissantz, Susanne. 2017. "Forming Virtual Teams – Visualization with Digital Whiteboards to Increase Shared Understanding, Satisfaction and Perceived Effectiveness" ICIS 2017 Proceedings. 9. https://aisel.aisnet.org/icis2017/SocialMedia/Presentations/9

Singh, Ronal Rajneshwar. 2018. "Designing for Multi-Agent Collaboration: A Shared Mental Model Perspective." https://minerva-access.unimelb.edu.au/items/5f5ea395-d150-51c7-980b-01b9321ac953

Sitterley, T. E. and Berge, W. A. Degradation of learned skills: Effectiveness of practice methods on simulated space flight skill retention. Seattle: The Boeing Company, Report 0180-15081-1, Prepared under NASA Contract No. NAS9-10962, 1972. https://ntrs.nasa.gov/citations/19730001426

Sørmo, Frode, Jörg Cassens, and Agnar Aamodt. 2005. "Explanation in Case-Based Reasoning–Perspectives and Goals." *Artificial Intelligence Review* 24 (2): 109–143. doi:10.1007/s10462-005-4607-7.

Stein, Dan J. 1992. "Schemas in the Cognitive and Clinical Sciences: An Integrative Construct." *Journal of Psychotherapy Integration* 2 (1): 45–63. doi:10.1037/h0101236.

Stone, Debbie, Jarrett, Caroline, Woodroffe, Mark and Minocha, Shailey 2005. User Interface Design and Evaluation. Morgan Kaufmann Series in Interactive Technologies. San Francisco: Morgan Kaufman.

Stout R. J., J. A. Cannon-Bowers, and E. Salas. 2017. "The role of shared mental models in developing team situational awareness: Implications for training." In *Situational Awareness,* Routledge, pp. 287–318. doi:10.4324/9781315087924-18

Sun, Liting, Wei Zhan, and Masayoshi Tomizuka. 2018. "Probabilistic Prediction of Interactive Driving Behavior via Hierarchical Inverse Reinforcement Learning." In 2018 21st International Conference on Intelligent Transportation Systems (ITSC), 2111–2117. doi:10.1109/ITSC.2018.8569453.

Sutcliffe, Alistair, and Uma Patel. 1996. "3D or Not 3D: Is It Nobler in the Mind?" In *People and Computers XI*, edited by Martina Angela Sasse, R. Jim Cunningham, and Russel L. Winder, 79–94. London: Springer London..

Sutcliffe, Alistair, Mark Ennis, and Jhen-Jia Hu. 2000. "Evaluating the Effectiveness of Visual User Interfaces for Information Retrieval." *International Journal of Human-Computer Studies* 53 (5): 741–763. doi:10.1006/ijhc.2000.0416.

Swaab, Roderick I., Tom Postmes, Peter Neijens, Marius H. Kiers, and Adrie C. M. Dumay. 2002. "Multiparty Negotiation Support: The Role of Visualization's Influence on the Development of Shared Mental Models." *Journal of Management Information Systems* 19 (1): 129–150. doi:10.1080/07421222.2002.11045708.

Van Bussel, F. J. 1980. "Human Prediction of Time Series." *IEEE Transactions on Systems, Man, & Cybernetics* 10 (7): 410–414. doi:10.1109/TSMC.1980.4308524.

Van den Bossche, Piet, Wim Gijselaers, Mien Segers, Geert Woltjer, and Paul Kirschner. 2011. "Team Learning: building Shared Mental Models." *Instructional Science* 39 (3): 283–301. doi:10.1007/s11251-010-9128-3.

Van Heusden, Arnold R. 1980. "Human Prediction of Third-Order Autoregressive Time Series." *IEEE Transactions on Systems, Man, & Cybernetics* 10 (1): 38–43. doi:10.1109/TSMC.1980.4308350.

Veldhuyzen, Wim, and Henk G. Stassen. 1977. "The Internal Model Concept: An Application to Modeling Human Control of Large Ships." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 19 (4): 367–380. doi:10.1177/001872087701900405.

Volz, K. 2018. "Cognitive Skill Degradation: Analysis and Evaluation in Flight Planning." (2018)./paper/Cognitive-skill-degradation%3A-Analysis-and-in-flight-Volz/134f87f2b5e5c1ddb756a8912481d69264b1afcc, https://dr.lib.iastate.edu/entities/publication/e7516e4a-e770-433c-989c-1f7794735f2a

Walsh, Sarah E., and Karen M. Feigh. 2021. "Differentiating 'Human in the Loop' Decision Strategies." In *IEEE Conference on Systems Man and Cybernetics*. Melbourn, Australia: IEEE. (virtual).

Wang, L., Z. Lin, H. Wang, and S. Chen. 2013. "Study on Motivation Mechanism of R D Team Creativity Based on Team Shared Mental Model." In 2013 Proceedings of PICMET '13: Technology Management in the IT-Driven Services (PICMET), 635–640. ISSN: 2159–5100.

Webber, Sheila Simsarian, Gilad Chen, Stephanie C. Payne, Sean M. Marsh, and Stephen J. Zaccaro. 2000. "Enhancing Team Mental Model Measurement with Performance Appraisal Practices." *Organizational Research Methods* 3 (4): 307–322. doi:10.1177/109442810034001.

Wegner, Daniel M. 1987. "Transactive Memory: A Contemporary Analysis of the Group Mind." In *Theories of Group Behavior*, edited by Brian Mullen and George R. Goethals, 185–208. New York, NY: Springer. doi:10.1007/978-1-4612-4634-3_9.

Wickens, Christopher D. 2008. "Situation Awareness: Review of Mica Endsley's 1995 Articles on Situation Awareness Theory and Measurement." *Human Factors* 50 (3): 397–403. doi:10.1518/001872008X288420.

Wood, David, Jerome S. Bruner, and Gail Ross. 1976. "The Role of Tutoring in Problem Solving." *Journal of Child Psychology and Psychiatry, and Allied Disciplines* 17 (2): 89–100. doi:10.1111/j.1469-7610.1976.tb00381.x.

Ryunosuke Yokoya, Tetsuya Ogata, Jun Tani, Kazunori Komatani and H. G. Okuno. 2007. "Discovery of other individuals by projecting a self-model through imitation." *IEEE/RSJ International Conference on Intelligent Robots and Systems,* 2007, pp. 1009–1014. doi: 10.1109/IROS.2007.4399153. https://ieeexplore.ieee.org/document/4399153

Yusoff, Nor'ain Mohd, and Siti Salwah Salim. 2020. *Shared Mental Model Processing in Visualization Technologies: A Review of Fundamental Concepts and a Guide to Future Research in Human-Computer Interaction*, Engineering Psychology and Cognitive Ergonomics. Mental Workload, Human Physiology, and Human Energy: 17th International Conference, EPCE 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part IJul 2020. Berlin, Heidelberg: Springer. 238–256. doi:10.1007/978-3-030-49044-7_20.

Zeiler, Matthew D., and Rob Fergus. 2014. "Visualizing and Understanding Convolutional Networks." In *Computer Vision – ECCV 2014*, edited by David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, 818–833. Cham: Springer International Publishing.

Zhang, Yu. 2008. "Role-Based Shared Mental Models." In 2008 International Symposium on Collaborative Technologies and Systems, 424–431. doi:10.1109/CTS.2008.4543960.

Ziebart, Brian D., Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. 2008. *Maximum Entropy Inverse Reinforcement Learning*, 6. Menlo Park, California: The AAAI Press. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence.