Received: 21 October 2022 | Revised: 29 September 2023 | Accepted: 9 October 2023

DOI: 10.1002/sys.21730

REGULAR ARTICLE

WILEY

A team-centric metric framework for testing and evaluation of human-machine teams

Jay Wilkins | David A. Sparrow | Caitlan A. Fealing | Brian D. Vickers | Kristina A. Ferguson | Heather Wojton

Operational Evaluation Division, Institute for Defense Analyses, Alexandria, Virginia, USA

Correspondence
Jay Wilkins, Operational Evaluation Division, Institute for Defense Analyses, Alexandria, Virginia, USA.
Email: lwilkins@ida.org

Abstract

We propose and present a parallelized metric framework for evaluating human-machine teams that draws upon current knowledge of human-systems interfacing and integration but is rooted in team-centric concepts. Humans and machines working together as a team involves interactions that will only increase in complexity as machines become more intelligent, capable teammates. Assessing such teams will require explicit focus on not just the human-machine interfacing but the full spectrum of interactions between and among agents. As opposed to focusing on isolated qualities, capabilities, and performance contributions of individual team members, the proposed framework emphasizes the collective team as the fundamental unit of analysis and the interactions of the team as the key evaluation targets, with individual human and machine metrics still vital but secondary. With teammate interaction as the organizing diagnostic concept, the resulting framework arrives at a parallel assessment of the humans and machines, analyzing their individual capabilities less with respect to purely human or machine qualities and more through the prism of contributions to the team as a whole. This treatment reflects the increased machine capabilities and will allow for continued relevance as machines develop to exercise more authority and responsibility. This framework allows for identification of features specific to human-machine teaming that influence team performance and efficiency, and it provides a basis for operationalizing in specific scenarios. Potential applications of this research include test and evaluation of complex systems that rely on human-system interaction, including—though not limited to—autonomous vehicles, command and control systems, and pilot control systems.

KEYWORDS
agent, artificial intelligence, human-machine team, human-system interaction, metric

1 | INTRODUCTION AND APPROACH

Increasingly capable human-machine teams (HMTs) will be a significant future component of industry, defense, medicine, and many other areas.[1–7] Increases in autonomous and learning capabilities in modern machines are driving the heightened importance of understanding and measuring HMT performance to ensure effective and safe collaboration with humans. These increased capabilities arise in the cognitive and physical realms alike, and the technologies employed by HMTs are advancing at a much more rapid pace than corresponding test and evaluation (T&E) concepts. This is particularly true of operational testing and evaluation (OT&E, or just OT) concepts, a situation which poses

significant challenges for decision makers in industry and defense, as systems already in the design or even developmental testing stages often lack both an accompanying, rigorous T&E program and a sufficient scope of capabilities and hazards for planning such a program.[8–13] Such pre-fielding test programs and policies for traditional technologies have been common for decades in both industry and defense, ensuring user safety and technological performance. Yet, as the technology outpaces our understanding of how to evaluate, utilize, and control it, much of that same rigor and policy guidance is lacking for scenarios where HMTs might play prominent roles.[12]

In this paper, we present a framework for categorizing and defining HMT performance assessment metrics, particularly those of interest to the T&E communities. The essential elements of this framework are the interactions between human(s) and machine(s) but also—and, to our purpose, equally if not more important—the functions of the team as a collective. As with any complex, multi-agent system, interactions between teammates or within any subgroup of teammates will introduce a new degree of unpredictability, potentially leading to propagations and emergent properties that assessments of individuals, alone, cannot capture.

For our purposes, a *teammate* of an agent is another agent[a]—human, machine, or otherwise—that (1) works in pursuit of the same goal(s) as the original agent, (2) has the ability to affect the current state of the relevant environment, system, or task, and (3) has the ability to coordinate action with the original agent. A *team* is a collection of agents, each of which is a teammate of all of the other agents within the team.[b] A *Human-Machine Team (HMT)* is a team consisting of at least one human agent and at least one machine agent guided by some type of computer or artificial intelligence. A human agent in an HMT is often referred to as an *operator*, and we occasionally use this terminology, also, to remain consistent with current literature. However, the reader should be careful not to conclude from this that the human is exclusively operating the machine agents; in that case, where the machines' actions are completely determined by a human controller, the machines would simply be *tools* and not teammates.

Since at least 2015, various commercial and government organizations have called for a new approach to Test, Evaluation, Verification, and Validation of autonomous and human-machine interactive systems.[14] The initial concern was the "state-space explosion," or the availability of so many courses of action and outcomes—some of which may not be explainable due to some of the black-box nature of artificial intelligence—that it becomes impossible to test a sufficient subset circumstances to provide assurance that the system will perform as desired or will not fail catastrophically in unanticipated ways.

Another concern was the intrinsic complexity of the agents' collective courses of action, themselves, which may be very difficult to plan for or replicate at the operational level and can make the prospect of obtaining assurances of performance and safety from testing even more daunting.[15] Self-driving cars, for instance, are currently among the most thoroughly tested of systems with some autonomous capabilities, yet there have been numerous fatal accidents. Initially, news media (and sometimes the legal system) often focus on who is at fault: the human or the machine?[16–18] However, almost all of these accidents are driven by failures in the human-machine teaming; see Stern's 2021 article, for instance.[19] Beyond the new approaches to testing required by autonomous systems, a new framework will be needed focusing primarily, if not specifically, on the human-machine teaming elements to enable both assessment and mitigation.

Our approach builds on and extends past efforts to categorize, define, and aggregate HMT assessment metrics in three key ways. First, we build a scalable framework that can be applied beyond the one human, one machine scenario, even up to teams of very large sizes, including potential machine swarms. Second, while individual human and machine metrics are vital and still must be included, we propose that the team, as a collective whole—how it is organized, how it interacts, how it adapts, and so on—should be the primary point of analysis, particularly if one is interested in the success and efficiency of team-based task completion. Lastly, we propose to break the tendency to focus almost exclusively on outcome-based assessments, which we claim are insufficient to support assessment of team effectiveness under a broad set of circumstances.

The second of the three preceding objectives is our primary guiding principle in developing this metric framework. As noted in our definition above, the key features of a team are coordinated action, in pursuit of a common goal, affecting the current state. Thus, if *team* performance or efficiency is not at a desired threshold, for instance, the root causes can often lie not exclusively or even primarily in the faults of any one individual agent but in the team's collective properties: its interaction and communication mechanisms; its collective cognitive and information transfer capabilities; its organization, architecture, and cohesion; its adaptability and ability to compensate for weaker agents, and so on. Evidence for this is common in assessments of human-only teams, and with little HMT operational T&E data available related to teaming metrics, we relied as much on the human teaming literature in this work as we did past human-machine teaming work.

In human-only teams, for instance, positive predictor-response and moderating relationships have been found between team performance and a variety of collective properties, including collective orientation—,that is, the tendency and willingness to work collectively in team scenarios—in complex problem solving scenarios,[20] and team cohesion as both a predictor of performance and a moderating factor between efficacy and performance.[21,22] Beyond outcome analysis, there is also evidence that other desirable team attributes are better understood with a collective perspective; Janssens et al., for instance, provide evidence that collective intelligence is best observed and measured not only through task outcomes but through the team's interactions, which, in turn, depend on the team's organization and information flow.[23]

Keeping the team as our central point of analysis, the framework we develop ultimately identifies the critical elements that characterize *teaming* and lays out categories for metrics as an extension of

[a] We use the language of agent-based modeling, and refer to all members of a team generically as *agents*.

[b] This is not to imply that *all* team agents—being teammates of all other agents—necessarily interact or coordinate. Our definition simply indicates that there is at least some pathway, however circuitous, for information to flow from any teammate to another, and the *potential* for coordination exists. Topologically, we are simply saying that a team—represented as a graph or simplicial complex—must be *connected*.

that. In many cases, we provide precise definitions of specific metrics and measures, either with our own, new definition or using a prior source, although some of our metrics are left broadly defined as classes of potential metrics focused on a narrower factor or set of factors. This breadth is intentional, as it is not our purpose in this paper to operationalize this metric framework towards specific T&E programs. Our objective is to provide an HMT metric framework that is specific enough to be of clear applicability to the T&E community, but that is broad enough to be a starting structure for identifying and selecting appropriate performance measures for a wide variety of teaming systems in T&E scenarios.

That said, operationalization is the next natural step, and one our team is already pursuing through formal T&E concepts with the US Department of Defense. Future articles will describe operationalization in general through simulations and eventually live tests, and we believe the framework we propose will prove quite robust. Indeed, measurements, observations, and data taken on the metrics that we discuss will provide insight not only into team performance and efficiency, but also what factors of the team led to that performance. In other words, our framework allows us to address whether a team is effective *in general*, not just during a particular observed task.

1.1 | What makes a human-machine team? Search and rescue as an illustration

To briefly illustrate a functional HMT and contrast it with a human simply using a machine tool, consider a search and rescue (S&R) mission performed by some number of humans and machines. A simple remote-controlled robot carrying medical supplies and food and water for the human would be an asset in S&R, but such a robot is merely a tool. Enabling the robot to follow the human does not change this very much, and it imposes minimal new requirements on testing and evaluation. A robot equipped with some search and reporting capability is currently feasible but, in the absence of coordinated action, is still in the "tool" category, even if equipped with the ability to measure heart rate, temperature, and so on. as part of a rudimentary triage. A robot, however, that can operate away from the human, decide on the basis of the triage whether to stay and attend to a victim or continue searching, can decide whether it is necessary to communicate this decision and when to do so, and can coordinate all of this information with other searching agents to optimize subsequent operations, is now in the *teammate* category. The ability of an autonomous system to operate at this level of complexity is or will very soon be within current capabilities.

At least three characteristics of the system—shared situational awareness, effectiveness of the interactions, and how the roles and responsibilities are shared or shifted between the teammates—lead to two changes in how we need to evaluate the system. First, the available state-space is now too large for comprehensive or timely exploration. Second, there are new measures and metrics beyond what is needed for testing a human with a tool.

2 | BACKGROUND AND MOTIVATION

There have been previous efforts to provide metric frameworks for HMT assessment, including comprehensive structures like ours that have identified and aggregated multiple performance metrics across a variety of categories. Damacharla et al.[24] and Singer and Akin[25] are good starting points. The latter is an excellent resource for performance and task completion metrics, and it also provides a good exposition of communication architecture and the building of metrics into HMT models. The former survey by Damacharla et al. goes a step further by including *teaming* metrics related to emergent, collective performance. Some of our framework derives from elements of Damacharla et al., but the teaming metrics described there are largely focused on outcomes and factors determined prior to assessment, such as mission assignment. There is little discussion of team interaction, interdependence, and their collective effects during the task.

Chauncey et al.[26] address co-adaptive HMT metrics, which assess changes in teammate behavior over time in response to environmental and interaction factors. They provide a detailed investigation of the human-machine interaction, but its focus is primarily on the two-agent scenario—one human, one machine. In 2008, Pina et al.[27] provided a very good discussion of general metric classes needed to assess HMT performance, using a conceptual behavior model as their base of analysis. They even discuss the multiple human-multiple robot cases and describe a case-study involving a single human-multiple robot search and rescue scenario. More recently, in 2017 Stowers et al.[31] developed a comprehensive metric framework based on a model of human, machine, and external inputs to human behavior and cognitive processes and states. This work was a particular inspiration to our own, especially in structure and derivation, but we wanted to remove the centralization of human processes and states and replace them with teaming characteristics.

Other notable works on metrics or metric frameworks include the 2018 MITRE guidebook,[28] which is motivated by defense applications and focuses on metrics in that direction; the 2003 work by Olsen and Goodrich,[29] which heavily focuses on the machine metrics, particularly those that measure the machine's autonomy with respect to a human operator; and the 2006 work by Steinfeld et al.,[30] which does devote discussion to the social aspects of human-machine teaming, but still only within the one human, one machine setting.

Our perspective in this proposed framework is to assess HMTs first and primarily as a *team*—that is, an interacting collective—with the characteristics of individual humans and machines seen (whenever possible) as instances of globally defined features of the team and its interactions. In both human-machine and human-only teaming, this approach is not new,[31–34] though it is still recognized as one that is difficult to utilize and requires further research and data. Yet, the number of researchers who, despite this difficulty, still take this team-centric point of view is a testament to the approach's importance.

This is true even for the development of the individual AI agents, themselves. Stowers et al.,[31] for instance, discuss team-level competencies needed for effective HMT performance, in the context of

surveying emerging AI technologies that could potentially bridge gaps in those competencies. Carroll et al.,[32] while they focus less on overall performance metrics and more on AI training methods for the machine agents in HMTs, provide data and evidence to support their claim that the best human-machine interaction and collaboration outcomes arise when the machine agents have been developed not in the usual AI paradigm of being trained on data in isolation or in competitive environments with other AI units, but when human data, even if suboptimal, is incorporated into the training model. Johnson and Vera[35] make the case in their 2019 article that a lack of teaming intelligence—an inability to effectively interact and collaborate with human teammates—is a significant and problematic gap in current intelligent systems, and one that can negatively impact both performance and robustness. They claim that a better understanding of the operations and capabilities of AI technology is incomplete without incorporating its potential as a teammate to human agents:

> In fact, technology thrives most when successfully woven into human work practice. Managing this integration requires teaming intelligence. (p. 18)

In the case of HMT modeling efforts, Miller et al.,[36] noted as recently as 2020 that current human-machine interaction modeling techniques do not readily support the design and assessment of systems that incorporate genuine, cooperative interaction between humans and AI agents, and that most such models still treat the human as a user external to the AI system. They then proceed, however, to take a significant step forward in bridging that gap by introducing a Systems Modeling Language (SysML) extension to structurally and dynamically model HMTs and their interactions within complex environments and through complex tasks. Their modeling effort included the ambitious goal to not only describe team interactions and performance but to provide a language for identifying requirements determined by interdependencies between the team agents. Ultimately, they show their modeling language to be capable of not only identifying interdependencies but describing team-level goals, team behavior, and particularly team structure—a component that our present framework also emphasizes as a teaming attribute that we believe has not yet been exploited to a fuller potential.

The aforementioned concept of interdependence—formally defined as the existence of complementary relationships between agents that are required to manage dependencies in joint activity[c]—is a crucial one in teaming research, and it shows up repeatedly as a fundamental mechanism for modeling interactions between team agents.[32,34–36] This makes it a concept that must be addressed in any attempt to model HMTs with the team as the centralizing component. Citing Johnson and Vera[35] again, for instance:

> Teamwork categories, characteristics, and properties vary from model to model, but the one concept that is consistent throughout is the importance of interdependence. This truth is both invariant across all domains and fundamental to teaming. … Understanding, supporting, and exploiting interdependence is what teaming intelligence is all about. (p. 18)

In their work on Coactive Design, Johnson et al.[37] use interdependence and joint activity as their central organizing design principles, thus placing the team at the center of their model for design of human-machine interactive systems. Even in the human-only teaming realm, Cooke and Gorman,[38] employ interdependence as the defining characteristic of a team of agents working together in their survey of methods for assessing team cognition at the collective level. Making interdependence the fundamental principle around which one constructs a design or assessment model is significant because—as we have emphasized in our own objectives—it places interaction between agents and, thus, the *team* at the origin, with all other relevant factors being extensions of the team's needs and performance. This is precisely the starting point for our framework.

Since our metric framework is directed towards performance assessment and not necessarily modeling of HMTs, we do not address interdependence directly. However, the concept is underlying nearly all of our team-level metrics, and certainly all of our team interaction metrics. Indeed, in our case, interaction can be taken as a synonym for interdependence.

3 | OUR APPROACH AND MOTIVATION

While our approach of making the team the central concept is not a new one, we do incorporate a novel parallel structure to the metric framework that functionally ties the human and machine agents together more as genuine teammates with common or comparable capabilities, as opposed to disparate agents with alien skill sets effectively using the other to accomplish a task. By starting with the team as a collective unit and assessing—from years of combined experience among the authors in operational testing and evaluation of human-systems integration and human-robot interaction in defense platforms—what interaction, performance, and efficiency factors were most important, we then extended and mapped those concepts onto individual traits of human and machine agents. This mapping revealed a sufficient correspondence to group the individual human and machine metrics into the same general categories. See Table 3 for a preview of this mapping, which shows that, in many cases, the correspondence between metric categories is one-to-one and even extends to some individual measures. To our minds, this parallel assessment structure between human and machine captured the literal spirit of teaming—bringing heterogeneous agents together to accomplish a task that could not have been carried out by any individual agent.

There are technological limitations that may prevent a full exploitation of this parallel HMT metric structure. There are still relevant

[c] In teaming language, an agent is in a state of—or possesses—*dependence* if that agent lacks the capacity required to competently perform a given activity in a particular context. An agent is *independent* if that agent possesses the capacity to perform a given activity in a particular context.

differences in machine and human cognition, and machines still do not have the capacity for features like emotional intelligence on par with humans, thus limiting the degree of perceived teammate likeness.[31] Nevertheless, we believe that such general AI is not a necessary prerequisite to examining HMT performance from this team-centric perspective, and even narrowly scoped HMT scenarios more than justify such an approach.

Additionally, while this metric framework is intended to apply across all sectors of application, it is also impossible for the authors to ignore the fact that our choices in this framework are colored by extensive experience in operational testing and evaluation, particularly of defense platforms and programs. Machine teammates in these scenarios are already expected to effectively interface with humans and make decisions based on data and human inputs, and requirements are very quickly reaching the stage where those machine agents will have to engage, make decisions, and choose courses of action that can have serious consequences and drastically and dangerously alter the environment, situation, and even the team, itself. For decisions and consequences of that magnitude, team trust and cohesion are critical. Consequently, our framework is implicitly conjecturing that the more parallels in goals and capabilities human-machine teammates observe between themselves, the more effective and better performing that team will be.

To assess HMTs through such a prism, we need metrics that focus on the *teaming* processes, which would especially include measurements of the organization of and interactions within the team, as well as traditional metrics on scenario outcomes. By collecting both interaction and outcome metric data, we can tie them together to better evaluate the factors that improve or deter overall team performance, or have more complex relationships with it. If these new metrics can capture the efficacy of the teaming, which we believe future test data will show, it may allow for extrapolation or generalization from the specific set of measurements performed.

3.1 | More than outcome-based testing

Another equally important motivating factor behind our development of this metric framework was our belief that operational testing and evaluation, especially in HMT scenarios, needs to move beyond considering exclusively *outcome-based testing* results. Outcome-based testing is simply a situation where a test scenario is generated, the test is carried out, and the *results* are evaluated, often only in relation to the factors that could be controlled or measured before the test; it is the most common form of operational testing in both defense and industry. However, the exclusive or primary focus on outcomes, along with controllable factors, tends to ignore the changes, interactions, reorganizations, developments, and so on. that can occur *during* the test. For a team, though, those in-test operations are the meat of the test; that is where the capabilities the team brings into the test are operated on by the team's learning capabilities and skill sets and translated into (hopefully successful) outcomes.

In defense of the testing community, these factors have been ignored largely because they are typically hard to capture and/or measure. Yet, they provide crucial insight; beyond simply determining if a test was successful or not, during-test factors can provide information as to why those particular outcomes occurred. So, if we can, somehow, measure those in-test factors, our knowledge and diagnostic abilities can only improve. Thankfully, as AI technology has advanced, so has the research and T&E communities' abilities to model and test complex phenomena.

Consider a version of our S&R example above. A remote-controlled robot carrying medical supplies would have the "usual" vehicle metrics to characterize performance: range, payload, communications reliability, and so on. This approach suffices when the requirements to be met are straightforward and separable, but human-machine teaming upends this paradigm with non-deterministic behavior over increasingly large state spaces. The usual metrics remain relevant, but must be supplemented by measures related to more complex behaviors. Enabling the robot to follow the human adds complexity via the autonomous navigation capabilities, but the coordination required by the team is straightforward, with well-established metrics. Going further to include course of action selection, suppose the robot decides, on the basis of the triage measurements, what course of action to take, and whether to communicate the decision to the human teammate. This creates the need for testing the coordination between the teammates in ways that allow for extrapolation to new scenarios. To allow for this extrapolation, we will need new metrics that can characterize not just the outcome but the quality of the communication and coordination—that is, the teaming.

In simple terms, the characteristics of the team determine what is possible for the team to accomplish, but we believe that now those characteristics should include the more complex, hard-to-measure interaction factors that play just as much of a role in performance as objective capabilities. Said another way, and referencing our discussion above on moving beyond just outcome-based testing, a team that performs well is not just one that accomplishes its task but one that can adapt and coordinate in order to accomplish those tasks in changing environments and under uncertain circumstances. Knowing that an HMT successfully accomplished a task or mission is not nearly as *diagnostically* valuable as knowing what particular attributes led to or hindered that outcome, particularly in operational T&E where a very small number of test events are used to extrapolate performance over very long periods of time in the future. The contributions, effectiveness, and robustness of the team's characteristics can be assessed in terms of the metrics in this framework and provide a deeper understanding of team potential that does not replace but enhances scenario-based outcome metrics. This perspective is central to our approach.

4 | HOW TO READ AND USE THIS FRAMEWORK

Our framework is a hierarchy of metric categories, sub-categories, classes of metrics, individual specific measures in some cases, and

definitions for all. The top-level consists exclusively of broad yet well-defined metric categories, and the second or mid-level consists of metric sub-categories, further decomposing the top-level by functional similarity. The final categorical level in this framework—what we call the *component* level—contains explicit measures in some cases and slightly more general metrics or classes of metrics in others. As we noted previously, our objective in this framework is to provide a broad foundation for HMT T&E assessment, but one that is still specific enough to provide guidance in in choosing appropriate assessment measures. In short, where we have not provided explicit, individual measures, we have at least provided a narrow enough class of metrics that a knowledgeable test developer or subject matter expert can determine a specific element of that class that is applicable to a given context.

The body of this article focuses primarily on the team-level metrics and measures, leaving much of, but not all of, the individual human and machine metrics discussion for the glossary in the appendix. We aim for this approach to be relevant to any mix of agents, including multiple humans, humans and machines, and larger groupings of humans and machines, including swarms. The parallelization of metrics across humans and machines is evident in the accompanying tables, as well as in our appended glossary of individual human and machine metrics, where many of the definitions in the team, human, and machine categories are functionally equivalent.

Since the individual metrics were chosen for their importance in supporting the performance assessment of the team, they were also chosen with an eye towards diagnosing when a particular problem with the team's decision making or performance is attributable to an individual teammate as opposed to shortcomings from the team as a whole. This assessment may be difficult to attribute to a single point (agent or team). In the self-driving car accident referred to earlier, the original failure attributed to the driver's failure to pay attention later shifted to the sensors' failure to identify the pedestrian in a timely fashion. However, this can also be seen as a teaming or concept of operations failure because of the unreasonable expectation of the human to quickly take control after a long period of boredom and inattention. Given that teaming failures can be accounted for and diagnosed at multiple places, failures can also be mitigated at these multiple places. It is then essential to determine whether the desired agent behavior is possible or likely and should be pursued; whether the teaming approach needs to be modified; or both.

Finally, while the metrics proposed here are doubtlessly incomplete and users will need to identify ones that are most useful for their system, we believe that this framework is still useful and practical. Our aim is to help system designers and evaluators begin thinking about the metrics that are relevant to their human-machine teams as they mature and increase in functionality. Metrics at team, human, and machine levels will, in general, be necessary for any analysis or diagnostics about what went wrong or what went right during a test scenario—diagnostics essential for designers and evaluators alike. In that sense, we hope this framework serves as a jumping-off point for tailoring key measurements and performance parameters to the team under evaluation.

TABLE 1 Top-level team metric categories and definitions.

Category	Definition
Capabilities	Abilities and capacities the team collectively possesses independent of any specific task, environment, or set of circumstances
Interaction	How team members engage in coordination, cooperation, and efficient goal pursuit during execution.
Performance	Qualitative and quantitative assessment of the team's decisions and actions, and the subsequent results and effects generated by or attributable collectively to the team.

5 | TEAM METRICS

Starting from high-level characterizations and progressing to finer detail, we established three broad categories of team-centered or collective metrics: *capabilities*, *interaction*, and *performance*, defined in Table 1. These originated from a view towards operational T&E, where evaluators generally take a somewhat chronological look at the evolution of a test event. To wit, we asked: (1) What objective, task-independent, measurable or identifiable attributes does a team possess *before* a task, or *bring into* a task? (2) What measurable or identifiable changes take place *during* a task? (3) What outcomes are identifiable and measurable *after the task* is complete? We defined the metric categories that answer these questions, respectively, as capabilities, interaction/interdependence[d], and performance.

Generally speaking, capabilities could range from qualitative properties like prior training and education a team is known to have, to numeric measures like average time to complete a particular task known from previous exercises, to things that might be known beforehand but not easily quantified, such as a team that is known from past experience to self-organize effectively in response to certain tasks, but for which the actual, resulting organization may not be revealed until the task is undertaken. A team's interaction metrics encompass the features of the team that are changed, used, or developed during the task, and that are enabled by the team's capabilities. Thus, for instance, if structure- and composition-based features are part of an HMT's capabilities—,for example, how the team is known to be organized or to self-organize, how it distributes roles and task allocations, and so on,—and if those capabilities are deemed relevant to a task, then its interaction metrics should include at least one or more metrics that measure the utilization of that structure and composition—that is, how the team utilized its organization and structure, how it communicated and allocated roles and tasks, how the agents engaged with each other and the environment, and so on. Lastly, the performance category focuses on a team's successes, failures, efficiency, and so on. with regard to outcomes and team decision making—,that is, assessments

[d] The metrics and metric classes in our team interaction category correspond to what most of the teaming literature—human-machine and human-only—refers to as interdependence. We have kept the terminology of interaction, feeling that it more aptly describes HMT actions during operational testing, but the important principle of interdependence in teaming underlies all of our interaction metrics.

TABLE 2 Top, second, and third-level teaming metrics.

Category	Sub-Category	Components
Capability	Communication framework	Information flow
		Joint world model
		Joint mission knowledge
	Synergy potential	Influence
		Role clarity
		Role adaptability
	Structure	Topology, hierarchical relationships, and other groupings
		Uni- & Bi-directional relationships
		Collective learning patterns
	Quantitative	Known numeric and/or statistical capabilities
Interaction	Team perspective	Situational awareness
		Information accuracy
	Cooperative behavior	Cohesion and goal management
		Intervention
		Team trust
		Agency shifting
	Team resource allocation	Joint attention allocation
		Workload transfer
		Endurance
Performance	Collective decision making	Optimality
		Robustness
		Risk level
	Collective task performance	Task success/failure
		Planning recognition
		Timeliness
		Efficiency

of results and effects attributable to the team's collective actions and decisions.

Table 2 shows the breakdown of the three top-level team metric categories into sub-categories (second-level) and then into the *component-level* metrics or metric classes.

5.1 | Team capabilities

The Capabilities category breaks down into four sub-categories: *Communication framework*, *synergy potential*, *structure*, and *quantitative*. The Quantitative category is purely for numerical or statistical capability attributes that do not fit into any of the other three categories. It might include measures computed from prior tests and assessments, like average time on a given task or mean time between operational failures for a network of machine agents.[e]

5.1.1 | Capability: Communication framework

Communication Framework refers to the timeliness, capacity, dynamics, precision, and accuracy of information transfer within the team; we also refer to this as *communication structure*, as it encompasses the means and processes by which the agents communicate, share, distribute, process, and assess information and knowledge. We identify three component metric classes under communication framework: *Information flow, joint mission knowledge, and joint world model*.

Information flow refers, intuitively, not only to what, but also when and how information is passed. Formally, information flow metrics encompass the mechanics and capacity for data and information transfer and retention among and between agents, including channels, directions, rates, control, interference, and system input-output.[39] By joint mission knowledge, we mean the compatibility or complementariness of agents' knowledge of the task or mission, including strategy for success and a basis for prioritizing competing goals or actions. The team members' knowledge need not be identical or even completely consistent, but it must enable collaboration for mutual support in pursuing the mission. Example areas likely to be key include red and blue team tactics, standard operating procedures, target weak points, and so on.[40] In a similar vein, joint world model refers to the extent to which agents in the team have compatible knowledge structures, meaning compatibility in what they can and cannot represent. As with joint mission knowledge, structures need not be alike (e.g., they may be complementary) but they must allow for collaboration with the other agent. See Damacharla et al.[24] and Madni & Madni[41] for further discussion. Note that in our framework, information flow depends upon compatible joint world models, which can be based on radically different teammate models and ontologies of the world due to current cognition differences between machines and humans.

5.1.2 | Capability: Synergy potential

We developed the Capability sub-category Synergy Potential as a way to capture those metrics that measure or describe the extent to which the team knows and understands how it is organized, each member's role and abilities to take on other roles, how the team might respond to the actions and choices of its members, and how the team can work together to optimally use all of those traits. It formally includes any and all aspects of the team architecture and agent interaction that

[e] Quantitative and statistical measures like these are generally obtained through prior T&E events. Thus, an outcome measurement in one test may very well end up being a capability input to another test.

define and support how team agents interact, are interdependent, and may influence one another. We define two specific metrics under synergy potential. *Influence* refers to the ability of one team agent to elicit a response—cognitive, verbal, emotional, behavioral, or otherwise—from other teammates.[42] *Role clarity and adaptability* encompass the awareness of, maintenance of, and—when necessary—adaptation of action boundaries (within the team and external to it), assigned roles and duties, action, and communication interfaces. Equivalently, it can refer to the awareness within the team to recognize roles and the potential need for role adjustments, and the flexibility and ability to implement necessary adjustments.[24,41] We expect that role clarity and adaptability (when allowable) will be of particular importance to HMT applications.

5.1.3 | Capability: Structure

The Structure metric sub-category refers to the team's topology or organizational shape, and includes potential factors such as divisions into and overlapping between subgroups within the team, whether the architecture is imposed and held fixed or organic and dynamic, hierarchies or flat relationships—that is, is there a chain of command directing the team, or is there an equal status between all the agents? As in traditional network models of teams, structure can also capture directed features—,for example, is there communication in only one direction between agents A and B, for instance, or can information flow between the two in either direction? We define three metrics under structure: *hierarchical relationships*, *learning patterns*, and *directional relationships*.

Hierarchical relationships encompasses the directional influence and command relationships within the team. For instance, if the team is modeled as a network with pairwise connections between agents, hierarchical relationships may weight or direct the edges connecting agents in that network model. Note that we treat hierarchy not as a structural property in and of itself, but as a qualitative one that affects structure—an understanding or knowledge of authority relationships between agents. This is because a team's underlying topology or structural organization (e.g., a network) need not necessarily be physically representative of any qualitative relationships between agents. Similarly, *Uni-* and *Multi-directional relationships* contain the characterizations of directional command, influence, communication, or action relationship between or among two or more team agents. We recognize that there is some conceptual overlap between Multi-Directional Relationships and Hierarchical Relationships, but each can exist without the other, which is why we include them separately. Lastly, *Learning Patterns* refers to those metrics that capture how the team agents' interactions work in conjunction with communication, information flow, and external sources to create collective team knowledge. Specifically, they are the method(s) by which the team absorbs, distributes, transfers, and assimilates information, as well as the method(s) by which the team's knowledge evolves over time in response to task, team architecture, and interaction.[43,44]

5.2 | Team interaction

Team interaction is how we broadly refer to the ways in which team members engage in coordination and cooperation during task and/or mission execution. As noted above, in our framework, interaction refers to the factors that the team controls and can change during a task. We identify three second-level metric categories under interaction: *Perspective*, *cooperative behavior*, and *team resource allocation*.

5.2.1 | Team interaction: Perspective

As a metric category, perspective captures what the team collectively knows and believes about the world, where we take "the world" to mean all aspects and circumstances external to, within, and among the team. As in the joint world model discussed above, this is taken as a joint definition, meaning that individual agents need not have identical perceptions of the world, but those perceptions should allow reinforcement and cooperation. Team perspective includes the collective situational awareness of the team. Given the differences in human and machine cognition, this will not necessarily be a completely shared (as in, commonly held) situational awareness—but must still support collective action. Similarly, the accuracy of the information passed across the team is a key element of the interaction, and will influence team performance. There are two components under perspective: *Information accuracy*, and *situational awareness*. Information accuracy is the quantification of correctness and precision of information passed between team members, as potentially measured against both absolute and relative standards.[39,45] Information accuracy can significantly affect metrics in other categories, including joint world and mission models, and its companion under the perspective heading—situational awareness.

Situational awareness is a widely studied concept in many different fields, including teaming, psychology, and cognition, and it is of particular importance in defense applications. There is extensive literature on individual and shared situational awareness in a teaming context, and it is closely related to concepts like memory and perception. For our framework, we allow situational awareness to be a class of potential metrics or measures at our component level, since the specific measurements are likely to be context-dependent. However, we do follow the widely-used general definition of situational awareness from Endsley,[46,47] taking it to be the extent to which the team perceives and distributes information regarding knowledge of collective group actions, tasks, or decisions based on the team's structure, interaction dynamics, comprehensive environment, relevant task/mission objectives, and effects, consequences, and influences of environmental factors. We further assume that situational awareness includes contextual awareness—the team's assessment of objective factors and their potential effects on task and its completion, including physical, social, cultural, economic, and political environments (i.e., the social terrain), friendly versus adversarial scenarios, and so on. It should also be noted that our view of assessing situational awareness at the team level for

HMTs closely follows that of Gorman, Cooke, et al.,[48] where situational awareness is assessed and understood through adaptation to task roadblocks by means of team coordination. In operational testing, this analog of task-based learning is often how teams develop *collective knowledge*, which—in addition to including team skills and "know-how" pertaining to particular tasks—is a broader umbrella term encompassing much of what we have already discussed in the team category, including joint world models, information accuracy, situational awareness, and so on.

5.2.2 | Team interaction: Cooperative behavior

The sub-category Cooperative Behavior comprises those metrics capturing the extent to which the team works together to attain a mutual goal or complementary goals. Cooperation both informs and follows from resource allocation (the third sub-category under Interaction), and management of goals is essential for effective cooperation. Thus, the quality and effectiveness of the cooperation must be assessed. We define four metrics under cooperative behavior: *agency shifting*, *cohesion*, *and goal management*, *intervention*, and *team trust*. Agency shifting describes how and when roles and control in the team change, the mechanisms by which those changes are carried out, as well as how the team adapts to filling different roles, having different levels of control, and changes in roles and control.[49]

Cohesion and intervention are related concepts that are both well-known in the human-robot interaction field. The former is the dynamic process that is reflected in the tendency of a group to remain united in the pursuit of its instrumental objectives and for the satisfaction of member needs. This is also where goal management is most essential; establishing clear objectives that the team believes it can accomplish together, making sure the team is aware of its intended tasks—including any dynamic changes in those objectives during execution, and maintaining collective and collaborative action towards those goals all are by-products and determining factors of team cohesion. Intervention is the extent to which teammates interact with and purposefully affect each other's actions. Interruption of or direct interaction in a teammate's actions may have a negative impact on overall team performance, but it may also be necessary to resolve errors or ensure that progress towards team goals remains on track. Previous research postulates that the relationship between intervention and performance is non-monotonic—an inverted *u*-shaped curve where some interactions improve performance but too many hurt it; see Damacharla et al.,[24] for instance, for further discussion.

Trust, generally speaking, is a difficult concept to define in any context. For HMT evaluation purposes, we take the sub-category team trust to contain any and all metrics pertaining to the collective and individual attitudes and beliefs that combined group effort will aid and succeed in achieving a team's goals, particularly in situations characterized by uncertainty and risk. Lee and See's three general bases of trust[50] in the context of automation—performance, process, and purpose—are very likely to play a role in HMTs. Based on previous trust in automation research, levels of trust in human-machine interactions may be ill-calibrated (i.e., overly trusting or insufficiently trusting) because of erroneous judgments of a teammate's performance, process, or purpose; see Madhavan & Weigmann[51] for further discussion. It is important to note that all of the cooperative behavior metrics depends upon internal trust within the team, which is directly tied to expectations, beliefs about teammates, and roles within the team.

5.2.3 | Team interaction: Resource allocation

Team resource allocation is the distribution and changing levels of cognitive, physical, and incentive factors among team members. It includes three component metrics or metric classes: *Endurance*, *joint attention allocation*, and *workload transfer*.

Endurance in our framework agrees with the commonly held notion of a team's collective capability to proceed on task under stress.[24,41] Joint attention allocation refers to measurements of the distribution of limited attention paid collectively to the team's tasks and the ability of the team to dynamically attend to and prioritize varied responsibilities such as strategic planning and current assignments. The descriptor "joint" in this term indicates that it describes the team's collective attention allocation, as opposed to an individual allocating his/her attention across multiple tasks. Workload transfer refers to the quantitative or qualitative description of any general process or mechanism by which one or more tasks is/are transferred from one subset of team members to another; this is similar to but more general than human-machine/machine-human intervention.[11,25]

A reasonably strong argument could be made, especially in operational testing scenarios, that team interaction both begins and ends with *resource allocation*—how the team allocates and redistributes resources and tasks. This is clearly a collective or interactive property, and it can have significant effects on other factors such as joint attention allocation, whereby the team must decide which team members attend to which elements of the mission. What aspects of a task are getting attention, from which teammates, and are there other important aspects that are not getting sufficient resources? In the S&R scenario, for example, if one agent is compelled to spend extra time treating and transporting a particular victim, it may be necessary for other agents to pick up the slack, covering more search ground and potentially treating more victims than originally planned. This could affect both the timing of the team's actions and decisions, as well as resources. If the structure permits, workload may be transferred among the team members, shifting in principle between humans and machines. There are endurance issues as well. This is not routinely considered for machines, but overheating, as an example, applies to machines as well as humans.

5.3 | Performance

Generally, the performance metric category captures those mission outcomes that will sound familiar to most readers, but it is important to remember that in this specific category, they refer to *collective team* outcomes. Specifically, we take the Performance category to contain all

qualitative and quantitative assessments of the team's decisions and actions, as well as the subsequent results and effects generated by or attributable collectively to the team. This is distinguished from any individual performance assessments and should only include measures that can be attributed to the team in some collective fashion. There are two Performance metric sub-categories: C*ollective decision making* and *collective task performance*.

5.3.1 | Performance: Collective task performance

Collective task performance encompasses all metrics capturing the quality of team productivity, efficiency, and other important outcome measures both for specific tasks and in general. A team may collectively succeed or fail at an overall task, even if some team members, respectively, fail or succeed at their individual subtasks, so the collective nature of these metrics is crucial. In particular, task or mission success is a collective measure and not a simple aggregate of individual successes. It will depend upon team capabilities and interactions.

We identify four metrics under collective task performance: *Efficiency*, *planning recognition*, *task success/failure*, and *timeliness*. Efficiency is what one would expect—those measures capturing how effectively the team performs while using the smallest amount of resources within its resource constraints. Planning recognition refers to the ability of the team to identify (one or more possible sets of) steps and subtasks necessary for the completion of the task as a whole.[52] Task success/failure refers to the traditional measure of a team's accuracy on a given task, regardless of efficiency and resource use. Success may be binary (good/bad, pass/fail, etc.), or it may be measured on a finer graded scale. Timeliness metrics capture the extent to which a task is completed both at and within a favorable or desirable time. This can be measured in different ways, often either as the absolute time to complete the task or a ratio of time-focused-on-task to time-assigned-to-task.

5.3.2 | Performance: Collective decision making

Formally, the collective decision-making sub-category captures those metrics that address the team's assessment of objective factors, options or choices of action, and their potential effects on the task and its completion potential. However, this decision making must be assessed directly—not indirectly through outcomes. A team might make suboptimal decisions but still have good or successful outcomes for a variety of reasons or sometimes out of pure chance, or they might make optimal decisions but still fall short of success because of, say, unavoidable environmental factors. The three component level metric classes under collective decision making are *optimality*, *risk level*, and *robustness*.

Optimality refers to the extent to which a decision maximizes utility and efficiency, or the extent to which the course of action taken maximizes the "value" of a decision, with value measured against a situationally appropriate standard. This could be something as straightforward as the expected outcome, if sufficient probability estimates can be assigned to the possible outcomes. [43,53] Notably, forming decisions according to statistical and mathematical models enabling optimal decision making is often complex, time-consuming, and requires more information than is usually available in real-life decision making.[53,54] Nevertheless, this is one area in which our group is already extending this framework, working on methods for determining optimal team structures for large HMTs (e.g., machine swarms).

Risk level assesses the extent to which the team's collective decisions or actions take account of the perceived threat of a given situation and the vulnerability expected from taking a certain action.[55] Human agents are generally poor judges of risk and probability, and yet are also generally risk averse, willing, for instance, to accept smaller expected outcomes with greater certainty than even slightly riskier alternatives with larger payoffs. This potential for contradiction means that risk assessment of HMT decision making should include some objective component.

Robustness includes any measure or observation of how effective the team's broad decision-making framework is at choosing action for different (types of) situations. Equivalently, and perhaps in more detail, robustness can be considered as the extent to which the collective decision-making process (1) accounts for the random and unpredictable nature of risk-inducing factors, and (2) yields decisions that are not just effective in typical or average scenarios—that is, have good *expected* performance—but also safe enough to utilize under a wide range of possible circumstances. Gigerenzer and Gaissmaier[56] have a useful variation of robustness, as well. Note that our definition of robustness is assumed to include measures or observations of adaptability and flexibility in decision making, including a team's ability to change or re-approach risk assessments and optimal outcomes during task execution.

6 | AGENT-LEVEL HUMAN AND MACHINE METRICS: PARALLEL T&E FUNCTIONALITY

We will not describe the individual human and machine agent metrics in the same detail as above. The discussion regarding our reasoning for choosing certain metrics follows very much as before, and the precise definitions (and sources, where applicable) are given in full in the appendix. Instead, we will briefly elaborate here on the parallel structure that arose in our metric framework between the human and machine individual metrics. This was not an expected outcome in developing our framework, but it did occur quite organically. We started only with two basic premises: placing the team—the collective unit—at the center of the assessment before looking at any individual agent metrics or measures (even before considering any distinction between human and machine agents), and making the framework useful for operational testing and evaluation but also in broader scenarios. Yet, after constructing the teaming metrics discussed in Section 5, and then extending outward to include individual metrics that were not just necessary for T&E but complimentary to the teaming metrics, we noted that there was a near one-to-one correspondence—in function, if not in name—between human and machine metrics.

Table 3 shows the human and machine metric top-level categories, the sub-categories, and the component level metrics or metric classes. The top-level metric categories are the same as in the team case. Capabilities refer to what qualities the human or machine agent brings to the team, independent of task, and Performance refers to the assessment of individual decision-making and outcome-generating actions that have impacts on the team's performance. Team interaction does not have as direct a counterpart at the individual agent level of analysis, though evaluation of the success and extent with which an agent cooperates with others to complete tasks should be carried out at the individual level. Thus, we instead focus on worldview because the agents' individual perspectives inform team interaction.

The human-level attributes chosen for this framework were the ones we deemed the most influential towards collective team performance. World view variables, for instance, account for how human agents think about and process tasks; when aggregated, distributed, or treated collectively holistically, these individual agent world-view metrics can provide team-level measures of attributes like situational awareness, attention allocation, and workload/fatigue.[38,48] Additionally, measuring and tracking performance variables at the individual human level allows for assessments of individual effects relating to one another in the context of a more global effect. For example, poor performance outcomes can be correlated with only a single member of a team lacking appropriate training and/or usability experience, even if the others have the necessary skills. Tracking human-level metrics also enables analyses linking to team-level and mission-level success more broadly (e.g., one member of a four-member team not relying on a communication system increases the likelihood of mission failure).

Individual machine capabilities also may not meet the requirements of the team, and be an underlying cause of failures in human-machine teaming. The machine platforms (e.g., robot, drone) will impose limitations on what the machine can contribute, potentially limiting team success. Additionally, the cognitive structures implemented in the software and hardware will impose constraints on what the human-machine team can undertake. Surprise or unexpected tasks arising within even a well-planned mission may be beyond the capabilities of the HMT.

The most prominent feature of the individual human and machine metrics in Table 3, however, is the parallel functionality between humans and machines. Reading across the table, along each row, shows the functional parallels between human and machine metrics. For instance, both humans and machines take in data from other agents and the environment, whether by means of sensory organs in the human case or signal sensors in the machine case.

The correspondences between human and machine categories and between human and machine sub-categories are directly one-to-one, and there are only a few places where the more refined component level metrics do not line up functionally. One such case on the human side is usability, which—by definition—refers to human assessments of usability of machine or mechanical systems. Generally speaking, machines do not utilize humans in the same way, at least not to the extent that scales like the System Usability Scale (SUS) have needed to be developed for machine use of human operators. Likewise, while intelligent machines can definitely have a constructed perspective or world-view, they do not yet possess emotional components that uniquely define human attitudes and perceptions. On the other hand, there are number of components to AI machine architecture, hardware, and software that do not have direct or explicit counterparts in human thought, cognition, and memory. There are peripherals, hardware and software components, rigid processing and operating margins, and platform restrictions for machine agents that humans can more resiliently overcome—,for example, if a human operator wants to completely overhaul a course of action and make a fundamental paradigm change in operating procedure, he or she can do that circumstances permitting, whereas a similar overhaul on the machine's part would involve completely changing its world model and value functions, an action that requires full retraining and is not yet a feasible capability of AI. In the other direction, machines are not plagued by the issues humans deal with in their innate (for the most part) inability to accurately and intuitively assess risk and probability.

Nevertheless, despite a few minor misalignments at the finest assessment levels, this parallel structure carries a number of advantages, in addition to being of potential academic interest. In particular, it supports our conjecture that many of the human-only teaming principles and results have (or will be shown to have) analogs for HMTs, allowing us to import existing work from the human teaming and system integration communities. The symmetry also allows for a wider set of topological structures and modeling relations between the human and machine.

7 | SUMMARY, CONCLUSIONS, AND FUTURE DIRECTIONS

Advances in artificial intelligence and autonomous systems allow machines to increasingly function as teammates with their human partners, rather than as tools used by the humans. As a teammate, the machine will share team goals, act to affect the state of affairs, and coordinate those actions with other teammates. In order to evaluate HMTs well, we especially need to focus on the interactions between agents and how agent-level attributes influence this interactive coordination.

The criticality of the human-machine interaction implies the need for an expanded assessment framework focused on metrics that capture the efficacy of the teaming per se. This paper describes such a framework, and presents an initial approach to the team-oriented metrics. The metrics addressing interaction span the categories of shared team perspective, cooperative behavior, and resource allocation within the team. These are the categories controlling the effectiveness of the teaming.

Our framework was developed with operational testing clearly in mind but not as a requirement, and the metrics and categories were chosen because of their impact on the teaming factors. A consequence of that team-centrality is that the framework parallelizes the individual human and machine metrics across functionality lines. The parallel functionality carries a number of advantages, easing import of existing

TABLE 3 Human and machine metric categories, sub-categories, and components. The parallel structure between human and machine metrics can be read across horizontal rows from the human sub-category (resp. component) column to the machine sub-category (resp. component) column.

Category	Human		Machine	
	Sub-categories	Components	Sub-categories	Components
Capability	Training and experience	Mental models	Cognition structures and algorithms	World model
		Mission knowledge		Mission knowledge
		Teammate knowledge and experience		Teammate knowledge and experience
	Psychological traits	Decisiveness and impulsiveness	Cognition hardware and software	Prioritization
		Flexibility		Algorithm flexibility
		Intelligence		Computation environment
	Physical abilities	Physical Fitness	Standard platform	Structural and mechanical elements
				Computer and peripherals
		Sensors (Organs/Equipment)		
				Integration hardware
Worldview	Judgments and attitudes	Situational awareness	Perspective	Situational awareness
		Trust		Trust
	Cognitive allocation	Working memory	Resource allocation	Process/Threat monitoring
		Attention allocation		
		Other dependability monitoring		Other dependability monitoring
	Effort	Resource availability (Workload, fatigue)	Resource use	Platform operating margin
		Usability		Processing
Performance	Decision making	Optimality	Decision making	Optimality
		Robustness		Robustness
		Risk level		Risk level
		Reliance		Reliance
	Task performance	Error rates	Task performance	Error rates
		Timeliness and efficiency		Timeliness and efficiency
		Other performance		Other performance

Note: Human and machine sub-categories appearing on the same row indicate that they are functionally aligned across humans and machines.

work from the human teaming and system integration communities and allowing for a broader set of topologies and hierarchical relationships when modeling structures for teams in T&E scenarios. We are already at a stage, for instance, where some machine safety features will override a human operator, and we will increasingly see situations where there is an interaction, but in the case of divergent views, the human will defer to the machine. Finally, the parallel structure supports generalization to larger teams.

Ongoing advances in computer hardware, artificial intelligence and autonomous systems will only continue to drive increasingly robust and complex human-machine teaming. Assessments of these developments in many fields will require a framework that focuses on the teaming itself. This will have implications for many if not all technologically oriented development communities. The various quality control, certification, Test & Evaluation, and Verification & Validation elements that provide both permissions and assurances for these developments will all be affected. Understanding, measuring, and impacting the performance of HMTs is essential to ensuring that the collaboration of humans and machines as teams can provide the safe and effective advantages that are expected of them. We offer this framework as another step on that path.

The most immediate future steps beyond this framework should focus first on operationalization, and then on verification of the feasibility and utility of these metrics through data. Toward that end, our group is in position to incorporate data collection into defense-based test concepts for new autonomous technologies, and we encourage other groups in commercial industry to do the same. Further on, we hope to build models and simulations of HMTs in various scenarios,

both for verification and validation as well as determination of team features needed for performance and efficiency optimization.

ACKNOWLEDGMENTS

The authors would like to thank The Institute for Defense Analyses for its financial and time support in the preparation of this article, and the reviewers for their very helpful suggestions.

DATA AVAILABILITY STATEMENT

This document has no accompanying data. Should any future data be associated with this manuscript and its claims, contents, and so on, the authors agree to make that data available upon request.

ORCID


Jay Wilkins https://orcid.org/0000-0002-7493-6210

REFERENCES

1. Clarke JP, Aerospace America. Human-machine teaming is key to the future of aerospace. 2020. Accessed Oct 20, 2022. Available at: https://aerospaceamerica.aiaa.org/year-in-review/human-machine-teaming-is-key-to-the-future-of-aerospace/
2. Konaev M, Chahal H, Brookings Institute *TechStream*. Building trust in human-machine teams. 2021. Accessed October 20, 2022. Available at: https://www.brookings.edu/techstream/building-trust-in-human-machine-teams/
3. Forbes CT, The importance of human-machine AI team-building. 2020. Accessed October 20, 2022. Available at: https://www.forbes.com/sites/forbestechcouncil/2020/10/05/the-importance-of-human-machine-ai-team-building
4. Daigle L, Military Embedded Systems. AI-enabled vehicles will usher in true human-machine teaming in the field. 2020. Accessed October 20, 2022. Available at: https://militaryembedded.com/ai/machine-learning/ai-enabled-vehicles-will-usher-in-true-human-machine-teaming-in-the-field
5. Kela-Medar N, Kela I. The machine-human collaboration in healthcare innovation. In: Brito SM, ed. *Toward Super-Creativity—Improving Creativity in Humans, Machines, and Human—Machine Collaborations*. IntechOpen. doi:10.5772/intechopen.88951
6. Lawless WF, Mittu R, Sofge D, Hiatt L. Artificial intelligence, autonomy, and human-machine teams—interdependence, context, and explainable AI. *AI Magazine*. 2019;40(3):5-13. doi:10.1609/aimag.v40i3.2866
7. Pellerin C, US Department of Defense. DoD News. Work: Human-machine teaming represents defense technology future. 2015. Accessed October 20, 2022. Available at: https://www.defense.gov/Explore/News/Article/Article/628154/work-human-machine-teaming-represents-defense-technology-future/
8. Insinna V, DefenseNews. Introducing Skyborg, your new AI wingman. 2019. Accessed February 23, 2023. Available at: https://www.defensenews.com/air/2019/03/14/introducing-skyborg-your-new-ai-wingman/
9. Insinna V, DefenseNews. Under Skyborg program, F-35 and F-15EX jets could control drone sidekicks. 2019. Accessed February 23, 2023. Available at: https://www.defensenews.com/air/2019/05/22/under-skyborg-program-f-35-and-f-15ex-jets-could-control-drone-sidekicks/
10. Lingel S, Hagen J, Hastings E, et al. *Joint All-Domain Command and Control for Modern Warfare: An Analytic Framework for Identifying and Developing Artificial Intelligence Applications*. RAND Corporation; 2020. Accessed February 23, 2023. Available in print and at https://www.rand.org/pubs/research_reports/RR4408z1.html
11. Demarest C, DefenseNews. How the Pentagon knows when JADC2 innovations are working. 2022. Accessed Feb 23, 2023. Available at: https://www.defensenews.com/battlefield-tech/it-networks/2022/11/17/how-the-pentagon-knows-when-jadc2-innovations-are-working/
12. Grady J, US Naval Institute (USNI) News. Panel: Pentagon needs to be clearer on goals for JADC2. 2022. Accessed February 21, 2023. Available at: https://news.usni.org/2022/04/06/panel-pentagon-needs-to-be-clearer-on-goals-for-jadc2
13. Hoehn JR, Joint All-Domain Command and Control (JADC2). Congressional research service IF11493; 2022. Accessed February 21, 2023. Available at: https://crsreports.congress.gov/product/pdf/IF/IF11493
14. Defense Science Board. Defense Technical Information Center (DTIC). Final report of the defense science board summer study on autonomy. DTIC Report AD1017790; 2016. Accessed October 20, 2022. Available at: https://dsb.cto.mil/reports/2010s/DSBSS15.pdf
15. Mindell D. *Our Robots, Ourselves: Robotics and the Myths of Autonomy*. 2015.
16. Statt N, The Verge. Tesla crash involving autopilot prompts federal investigation. 2018. Accessed October 13, 2022. Available at: https://www.theverge.com/2018/5/16/17363158/nhtsa-tesla-autopilot-crash-Investigation
17. Lee TB, ArsTechnica. Uber self-driving car hits and kills pedestrian [Updated]. 2018. Accessed October 1, 2022. Available at: https://arstechnica.com/cars/2018/03/uber-self-driving-car-hits-and-kills-pedestrian/
18. Lee TB, ArsTechnica. Report: Software bug led to death in Uber's self-driving crash. 2018. Accessed August 1, 2021. Available at: https://arstechnica.com/techpolicy/2018/05/report-software-bug-led-to-death-in-ubers-self-driving-crash
19. Stern R, Phoenix New Times. Was the backup driver in an Uber autonomous crash wrongfully charged? 2021. Accessed October 16, 2021. Available at: https://www.phoenixnewtimes.com/news/uber-self-driving-crash-arizona-vasquez-wrongfully-charged-motion-11583771
20. Hagemann V, Kluge A. Complex problem solving in teams: the impact of collective orientation on team process demands. *Front Psychol*. 2017;8:1730. doi:10.3389/fpsyg.2017.01730
21. Beal DJ, Cohen RR, Burke MJ, Mclendon CL. Cohesion and performance in groups: a meta-analytic clarification of construct relations. *J App Psych*. 2003;88(6):989-1004.
22. Mathieu JE, Kukenberger MR, D'Innocenzo L, Reilly G. Modeling reciprocal team cohesion-performance relationships, as impacted by shared leadership and members' competence. *J App Psych*. 2015;100(3):713-734. https://doi.org/10.1037/a0038898
23. Janssens M, Meslec N, Leenders RThAJ. Collective intelligence in teams: contextualizing collective intelligent behavior over time. *Front Psychol*. 2022;13:989572. doi:10.3389/fpsyg.2022.989572
24. Damacharla P, Javaid AY, Gallimore JJ, Devabhaktuni VK. Common metrics to benchmark human-machine teams (HMT): a review. *IEEE Access*. 2018;6:38637-38655. doi:10.1109/ACCESS.2018.2853560
25. Singer S, Akin D, A survey of quantitative team performance metrics for human-robot collaboration. *41st Int. Conf. on Environmental Systems* 2012 (published online). doi:10.2514/6.2011-5248
26. Chauncey K, Harriott C, Prasov Z, Cunha M, A Framework for Co-adaptive Human-Robot Interaction Metrics. *Proc. Workshop on Human-Robot Collaboration: Towards Co-Adaptive Learning Through Semi-Autonomy and Shared Control* IEEE/RSJ International Conference on Intelligent Robots and Systems. 2016;1-5.
27. Pina PE, Cummings ML, Crandall JW, Penna MD, Identifying generalizable metric classes to evaluate human-robot teams. In *Proceedings of the 3rd Annual Conference on Human–Robot Interaction*, New York, NY: ACM; 2008;13-20.

28. McDermott P, Dominguez C, Kasdaglis N, Ryan M, Trahan I, Nelson A, Human-Machine Teaming Systems Engineering Guide. MITRE Product MP180941. The MITRE Corporation; 2018.
29. Olsen DR, Goodrich M. Metrics for evaluating human-robot interaction. *Performance Metrics for Intelligent Systems (PerMIS) Workshop*. 2003.
30. Steinfeld A, Fong T, Kaber D, et al, Common metrics for human-robot interaction. *Proc. 1st ACM SIGCHI/SIGART Conf. Human-Robot Interaction*. 2006:33-40. doi:10.1145/1121241.1121249
31. Stowers K, Brady LL, Maclellan C, Wohleber R, Salas E. Improving team competencies in human-machine teams: perspectives from team science. *Front Psychol*. 2021;12:590290. doi:10.3389/fpsyg.2021.590290
32. Carroll M, Shah R, Ho MK, et al. On the utility of learning about humans for human-AI coordination. *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems* 2019. Eds. Wallach HM, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox EA, Garnett R. December 8–14, 2019; Vancouver, BC, Canada, 5175-5186.
33. Seeber I, Bittner E, Briggs RO, et al. Machines as teammates: a research agenda on AI in team collaboration. *Inf Manag*. 2020;57:103174. doi:10.1016/j.im.2019.103174
34. Stowers K, Oglesby J, Sonesh S, Leyva K, Iwig C, Salas E. A framework to guide the assessment of human-machine systems. *Hum. Factors*. 2017;59:172-188. doi:10.1177/0018720817695077
35. Johnson M, Vera AH. No AI is an Island: the case for teaming intelligence. *AI Mag*. 2019;40(1):16-28.
36. Miller ME, Mcguirl JM, Schneider MF, Ford TC. Systems modeling language extension to support modeling of human-agent teams. *Syst Eng*. 2020;23:519-533.
37. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, Van Riemsdijk MB, Sierhuis M. Coactive design: designing support for interdependence in joint activity. *J Hum Robot Interact*. 2014;3(1):43-69. doi:10.5898/JHRI.3.1.Johnson
38. Cooke NJ, Gorman JC. Assessment of team cognition. In: Karwowski W, ed. *International Encyclopedia of Ergonomics and Human Factors*. 2nd ed. Taylor and Francis Ltd; 2006:270-275.
39. Arndt C. Information Measures: Information and its Description in Science and Engineering *Berlin*. Springer; 2001.
40. Canonico L, All Dissertations (Clemson University). Human-machine teamwork: an exploration of multi-agent systems, team cognition, and collective intelligence. 2019. Accessed June 12, 2022. Available at: https://tigerprints.clemson.edu/all_dissertations/2490
41. Madni A, Madni C. Architectural framework for exploring adaptive human-machine teaming options in simulated dynamic environments. *Systems*. 2018;6(4):44. doi:10.3390/systems6040044
42. Crichton MT, Flin R. Identifying and training non-technical skills of nuclear emergency response teams. *Ann. Nucl. Energy*. 2004;31(12):1317-1330. doi:10.1016/j.anucene.2004.03.011
43. Greening BR, Pinter-Wollman N, Fefferman NH. Higher-Order Interactions: understanding the knowledge capacity of social groups using simplicial sets. *Curren Zool*. 2015;61(1):114-127. doi:10.1093/czoolo/61.1.114
44. Greening Jr BR, Rutgers University dissertation. Higher-order analysis of knowledge capacity and learning potential in social animal groups. Accessed September 22, 2021. Available at: https://rucore.libraries.rutgers.edu/rutgers-lib/45285/ doi:10.7282/T3RB7325
45. Burnham K, Anderson D. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. Springer; 2002.
46. Endsley MR. Measurement of situation awareness in dynamic systems. *Hum. Factors*. 1995;37(1):65-84. doi:10.1518/001872095779049499
47. Endsley MR. Theoretical underpinnings of situation awareness: a critical review. In: Endsley MR, Garland DJ, eds. *Situation Awareness Analysis and Measurement*. Lawrence Erlbaum Associates; 2000:3-32.
48. Gorman JC, Cooke NJ, Pederson HK, Connor OO, DeJoode JA, Coordinated awareness of situation by teams (CAST): measuring team situation awareness of a communication glitch. *Proc. Hum. Factors and Erg. Soc. 49th Annual Meeting*. 2005;274-277.
49. Derue DS, Morgeson FP. Stability and change in person-team and person-role fit over time: the effects of growth satisfaction, performance, and general self-efficacy. *J Appl Psychol*. 2007;92(5):1242-1253. doi:10.1037/0021-9010.92.5.1242
50. Lee JD, See KA. Trust in Automation: designing for Appropriate Reliance. *Hum. Factors*. 2004;46(1):50-80. doi:10.1518/hfes.46.1.50_30392
51. Madhavan P, Wiegmann DA, A new look at the dynamics of human-automation trust: is trust in humans comparable to trust in machines? *Proc. of the Human Factors and Ergonomics Society Annual Meeting* 2004;48(3):581-585. doi:10.1177/154193120404800365
52. Bolia RS, Nelson WT, Summers SH, et al, Collaborative decision making in network-centric military operations. *Proc. of the Human Factors and Ergonomics Society Annual Meeting* 2006;50(3):284-288. doi:10.1177/154193120605000316
53. Becker GM, Mcclintock CG. Value: behavioral decision theory. *Annu. Rev. Psychol*. 1967;18(1):239-286. doi:10.1146/annurev.ps.18.020167.001323
54. Kurz-Milcke E, Gigerenzer G. Heuristic decision making. *Marketing: J Res Manag*. 2007;3(1):48-56.
55. Krokhmal P, Murphey R, Pardalos P, Uryasev S, Zrazhevski G. Robust decision making: addressing uncertainties in distributions. *Cooperative Control: Models, Applications and Algorithms (Vol 1)*. 2003. doi:10.1007/978-1-4757-3758-5_9
56. Gigerenzer G, Gaissmaier W. Decision Making: nonrational Theories. In: Smelser NJ, Baltes B, eds. *International Encyclopedia of the Social & Behavioral Sciences*. Elsevier; 2015. doi:10.1016/B978-0-08-097086-8.26017-0
57. Laird J. *The Soar Cognitive Architecture*. MIT Press; 2012.
58. Buchanan BG, Davis R, Smith RG, Feigenbaum EA. Expert systems: a perspective from computer science. In: Ericsson KA, Hoffman RR, Kozbelt A, Williams AM, eds. *The Cambridge Handbook of Expertise and Expert Performance*. Cambridge University Press; 2018. doi:10.1017/9781316480748.007
59. Samek W, Montavon G, Vedaldi A, Hansen LK, Müller K-R. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer International Publishing; 2019. doi:10.1007/978-3-030-28954-6
60. PN J-L. Mental models. Foundations of cognitive science. *Posner MI*. MIT Press; 1989.
61. Staggers N, Norcio AF. Mental models: concepts for human–computer interaction research. *Int J Man–Mach Studies*. 1993;38(4):587-605. doi:10.1006/imms.1993.1028
62. Kirk JR, Laird JE, Learning hierarchical symbolic representations to support interactive task learning and knowledge transfer. *Proceedings of the 28th International Joint Conference on Artificial Intelligence* 2019;6095-6102. doi: 10.24963/ijcai.2019/844
63. Mininger A, Laird JE, Using Domain Knowledge to Correct Anchoring Errors in a Cognitive Architecture. *Proceedings of the Seventh Annual Conference on Advances in Cognitive Systems* 2019:1-17.
64. Ramaraj P, Sahay S, Kumar SH, Lasecki WS, Laird JE. Towards using transparency mechanisms to build better mental models. *Advances in Cognitive Systems 7th Goal Reasoning Workshop*. 2019;7:1-6.
65. Department of The Army (USA)—US Army Training and Doctrine Command. Military Intelligence Company and Platoon Reference Guide. Publication No. TC 2–19.01. 2021. Accessed June 1, 2022. Available at: https://armypubs.army.mil/ProductMaps/PubForm/Details.aspx?PUB_ID=1021734
66. Anastasi A. Traits, states, and situations: a comprehensive view. In: Wainer H, Messick S, eds. *Principles of Modern Psychological Measurement*. 1983.

67. Sandini G, Metta G, Vernon D. The iCub cognitive humanoid robot: an open-system research platform for enactive cognition. In: Bongard J, Pfeiger R, eds. 50 Years of Artificial Intelligence: Essays Dedicated to the 50th Anniversary of Artificial Intelligence. Lecture Notes in Computer Science. *Lungarella M, Iida F*. Springer; 2007. doi:10.1007/978-3-540-77296-5_32
68. Kruglanski AW, Webster DM. Motivated closing of the mind: seizing and freezing. *Psychol. Rev*. 1996;103(2):263-283. doi:10.1037/0033-295X.103.2.263
69. Neuberg SL, Newsom JT. Personal need for structure: individual differences in the desire for Simpler structure. *J Pers Soc Psychol*. 1993;65(1):113-131. doi:10.1037/0022-3514.65.1.113
70. Roets A, Van Hiel A. Separating ability from need: clarifying the dimensional structure of the need for closure scale. *Personal Soc Psychol Bull*. 2007;33(2):266-280. doi:10.1177/0146167206294744
71. Weissman MS. Decisiveness and psychological adjustment. *J. Pers. Assess*. 1976;40(4):403-412. doi: 10.1207/s15327752jpa4004_10
72. Potworowski GA, Deep Blue Documents (Univ. Michigan Doctoral Dissertations). Varieties of indecisive experience: Explaining the tendency to not make timely and stable decisions. Accessed July 13, 2022. Available at: https://deepblue.lib.umich.edu/handle/2027.42/75906
73. Dickman SJ. Functional and dysfunctional impulsivity: personality and cognitive correlates. *J. Pers. Soc. Psychol*. 1990;58(1):95-102. doi:10.1037//0022-3514.58.1.95
74. Mackillop J, Weafer J, C Gray J, Oshri A, Palmer A, De Wit H. The latent structure of impulsivity: impulsive choice, impulsive action, and impulsive personality traits. *Psychopharmacology (Berl.)*. 2016;233(18):3361-3370. doi:10.1007/s00213-016-4372-0
75. Whiteside SP, Lynam DR. The Five Factor Model and impulsivity: using a structural model of personality to understand impulsivity. *Person. Indiv. Differ*. 2001;30(4):669-689. doi:10.1016/S0191-8869(00)00064-7
76. Wang J-J, Jing Y-Y, Zhang C-Fa. Optimization of capacity and operation for CCHP system by genetic algorithm. *Appl. Energy*. 2010;87(4):1325-1335. doi:10.1016/j.apenergy.2009.08.005
77. Bryson J. Cross-paradigm analysis of autonomous agent architecture. *J. Exp. Theor. Artif. Intell*. 2000;12(2):165-189. doi:10.1080/095281300409829
78. VandenBos GR. *APA Dictionary of Psychology*. American Psychological Association; 2007.
79. Martin AJ. Adaptability—what it is and what it is not: comment on Chandra and Leong (2016). *Am. Psychol*. 2017;72(7):696-698. doi:10.1037/amp0000163
80. Martin AJ, Nejad H, Colmar S, Liem GAD. Adaptability: conceptual and empirical perspectives on Responses to change, novelty, and uncertainty. *Aust J Rehabil Couns*. 2012;22(1):58-81. doi:10.1017/jgc.2012.8
81. Mcgrew KS. CHC theory and the human cognitive abilities project: standing on the shoulders of the giants of psychometric intelligence research. *Intelligence*. 2009;37(1):1-10. doi:10.1016/j.intell.2008.08.004
82. Sternberg RJ. The theory of successful intelligence. *Interam J Psychol*. 2005;39(2):189-202.
83. Sternberg RJ. A theory of adaptive intelligence and its relation to general intelligence. *J Intell*. 2019;7(4):23. doi:10.3390/jintelligence7040023
84. Raven JC, Court JH, Raven JE. In: Lewis H K, ed. *Manual for Raven's Progressive Matrices and Vocabulary Scales, Rev*. 1988.
85. Monsell S. In: Bruce V, ed. *Unsolved Mysteries of the Mind: Tutorial Essays in Cognition*. 1996:93-148. Control of mental processes.
86. Hogan J. Structure of physical performance in occupational tasks. *J. Appl. Psychol*. 1991;76(4):495-507. doi:10.1037/0021-9010.76.4.495
87. Fleming AJ, Behrens S, Moheimani SOR. Innovations in piezoelectric shunt damping. *Smart Struct Devices*. 2001;4235:89-101. doi:10.1117/12.420891
88. Whitten J, Bentley L. *System Analysis and Design Methods*. 7th ed. 2007.
89. Hastie R. Problems for judgment and decision making. *Annu. Rev. Psychol*. 2001;52(1):653-683. doi:10.1146/annurev.psych.52.1.653
90. Siciliano B, Khatib O. *Springer Handbook of Robotics*. Springer-Verlag; 2008.
91. Jajodia S, Liu P, Coyan V, Wang C. *Cyber Situational Awareness*. Springer; 2010.
92. National Institute of Standards and Technology. The Five Functions. Cybersecurity Framework. Online Learning Module; 2018. Accessed October 1, 2022. Available at: https://www.nist.gov/cyberframework/online-learning/five-functions
93. Parasuraman R, Sheridan TB, Wickens CD. Situation awareness, mental workload, and trust in automation: viable, empirically supported cognitive engineering constructs. *J Cogn Eng Decis Mak*. 2008;2(2):140-160. doi:10.1518/155534308x284417
94. Wilson BG, Cole P. In: Jonassen DH, ed. *Handbook of Research in Instructional Technology*. 1996:601-621. Cognitive Teaching Models.
95. Baddeley A. Working memory. *Curr. Biol*. 2010;20(4):R136-R140. doi:10.1016/j.cub.2009.12.014
96. Engle RW. Working memory capacity as executive attention. *Curr. Dir. Psychol. Sci*. 2002;11(1):19-23. doi:10.1111/1467-8721.00160
97. Conway ARA, Kane MJ, Bunting MF, Hambrick DZ, Wilhelm O, Engle RW. Working memory span tasks: a methodological review and user's guide. *Psychon. Bull. Rev*. 2005;12(5):769-786. doi:10.3758/BF03196772
98. Cowan N. *Working Memory Capacity: Classic Edition*. Psychology Press; 2016. doi:10.4324/9781315625560
99. Archibald LMD, Levee T, Olino T. Attention allocation: relationships to general working memory or specific language processing. *J. Exp. Child. Psychol*. 2015;139:83-98. doi:10.1016/j.jecp.2015.06.002
100. Lysaght RJ, Hill SG, Dick AO, Plamondon BD, Linton PM, Defense Technical Information Center Report: Operator Workload: Comprehensive Review and Evaluation of Operator Workload Methodologies. 1989. Accessed September 22, 2022. Available at: https://apps.dtic.mil/sti/citations/ADA212879
101. Cain B, Defense Research and Development Canada Report: A review of the mental workload literature. Report No. RTO-TR-HFM-121-Part-II. Toronto, Canada. 2007. Accessed via: USA Defense Technical Information Center. Accessed October 19, 2022. Available at: https://apps.dtic.mil/sti/citations/ADA474193
102. Hart SG, Nasa-Task Load Index (NASA-TLX); 20 Years Later. Proceedings of the Human Factors and Ergonomics Society Annual Meeting. 2006;50(9):904-908. doi:10.1177/154193120605000909
103. Hart SG, Staveland EL. Development of NASA-TLX (Task Load Index): results of empirical and theoretical research. *Advances in Psychology*. 1988;52:139-183. doi:10.1016/S0166-4115(08)62386-9
104. Hockey R. *The Psychology of Fatigue: Work, Effort and Control*. Cambridge University Press; 2013. doi:10.1017/CBO9781139015394
105. Nielsen J, Usability Inspection Methods. In: *Conference companion on Human factors in computing systems*. Association for Computing Machinery. 1994:413-414. doi:10.1145/259963.260531
106. Brooke J. SUS: a quick and dirty usability scale. In: Jordan P, Thomas B, Weerdmeester B, McClelland L, eds. Usability Evaluation in Industry. 1996:189-194. doi:10.1201/9781498710411-35
107. Veres SM, Molnar L, Lincoln NK, Morice CP. Autonomous vehicle control systems — a review of decision making. *Proceedings of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering*. 2011;225(2):155-195. doi:10.1177/2041304110394727

108. Kahneman D, Tversky A. Prospect theory: an analysis of decision under risk. *Econometrica*. 1979;47(2):263-291. doi:10.2307/1914185
109. Madhavan P, Wiegmann DA. Similarities and differences between human–human and human–Automation trust: an integrative review. *Theor Issues Ergon Sci*. 2007;8(4):277-301. doi:10.1080/14639220500337708
110. Borman WC, Motowidlo SJ. Task performance and contextual performance: the meaning for personnel selection research. *human performance*. 1997;10(2):99-109. doi:10.1207/s15327043hup1002_3
111. Cothier PH, Levis AH. Timeliness and measures of effectiveness in command and control. *IEEE Trans. Syst. Man Cybern.* 1986;16(6):844-853. doi:10.1109/TSMC.1986.4309003
112. Locke EA, Shaw KN, Saari LM, Latham GP. Goal setting and task performance: 1969–1980. *Psychol. Bull.* 1981;90(1):125-152. doi:10.1037/0033-2909.90.1.125
113. Maddi SR, Matthews MD, Kelly DR, Villarreal B, White M. The role of hardiness and grit in predicting performance and retention of USMA cadets. *Mil Psychol*. 2012;24(1):19-28. doi:10.1080/08995605.2012.639672
114. Lucas GM, Gratch J, Cheng L, Marsella S. When the going gets tough: grit predicts costly perseverance. *J. Res. Personal.* 2015;59:15-22. doi:10.1016/j.jrp.2015.08.004

How to cite this article: Wilkins J, Sparrow DA, Fealing CA, Vickers BD, Ferguson KA, Wojton H. A team-centric metric framework for testing and evaluation of human-machine teams. *Systems Engineering*. 2024;27:466–484. https://doi.org/10.1002/sys.21730

APPENDIX: HUMAN AND MACHINE METRIC VOCABULARY AND DEFINITIONS

Remark: *Top-level metric categories are fully left-justified. Metric subcategories (2nd level) are slightly indented, and component-level metrics or metric classes are the most indented. Metrics or metric categories that are functionally aligned between humans and machines—that is, parallel, in our framework terminology—are grouped together*.

Capability (Human and Machine): Human agent physical and psychological abilities and capacities—or machine agent operating and sustaining abilities and capacities—that are independent of any specific task, environment, or circumstances (but still relevant to any environment, task, or application the team might engage).

Training and Experience (Human): Measurable or identifiable indicators of a human operator's psychological and physical readiness for a task, including knowledge of the task, knowledge of relevant systems, natural talent, and prior interaction with other agents and systems that impacts the operator's ability to effectively perform the mission.

Cognition Structures & Algorithms (Machine): The knowledge structures and processes that the machine uses to represent information about the world, mission, and teammate(s). These capabilities inform task selection, task accomplishment, teaming, and communication, among related performance metrics. While there are many ways to computationally represent and store the information to be used by the machine,[57,58] and while these systems and knowledge types may or may not be explicit or "explainable" in the sense of Samek,[59] during T&E, critical capabilities must be testable.

Mental Model (Human): Representations of knowledge (the world, teammates, objects, processes) that can be manipulated for reasoning processes.[60] For T&E purposes, this can only be imperfectly estimated, with different communities using different terms, as discussed, for instance, by Staggers & Norcio.[61]

World Model (Machine): The structure for the machine knowledge, or what it can and cannot represent.[62–64] A key point is that the structure is not required to be complete or high fidelity; it must only be sufficient to support collaboration with any teammate(s) operating with a similar and/or different worldview.

Task/Mission Knowledge (Human[65] and Machine[62–64]): Similar to joint mission knowledge for the team, with a change in scope to the individual; an agent's information about and understanding of the task or mission. Consists of representations within a model allowing for flexible mission accomplishment, including strategies for success, bases for prioritizing competing goals or actions, and so on.

Teammate Knowledge and Experience (Human): The range of the human agents' understanding of or information on the machine teammate(s), including operational guidelines and capabilities, specifications, troubleshooting routines, and the humans' ability or proficiency with the system(s), acquired through previous training or practice.

Teammate Knowledge and Experience (Machine): Aggregation of representations allowing for the machine to coordinate activities with any teammate(s). These include up the agent's training and skills, orientation, best and most appropriate communication method(s), and so on. Potentially, this could be updated during execution. This also includes the machine's ability or proficiency with the human(s).

Psychological Traits (Human[66]): Enduring or frequently called-upon cognitive and behavioral characteristics that differentiate an individual human agent from others. Measurable traits are expected to be manifested in the observable behavior of individuals, keeping in mind that situation, state, environment, and task can all affect behavior.

Cognition Software and Hardware (Machine): Software that executes or supports higher-level capabilities (Cognitive Algorithms). Cognition software includes[67] processes analogous to the lower-level human cognitive functions that allow processing to occur—,for example, operating system, firmware, and other basic processes for execution of computational processing. Controller and interface functions may be separated into software and hardware components during developmental testing and evaluation, but would typically not be separated during operational testing and evaluation.

Decisiveness and Impulsiveness (Human): Decisiveness is the tendency to engage—or the momentum with which one

engages—in the decision-making process, particularly a tendency to simplify information and, thus, the decision making, regardless of the quality of the decision.[68–71] As Potworowski[72] has noted, decisiveness can be generalized or context-/domain-specific, so selection of the appropriate scale will need to be based on the researcher's intent. Impulsiveness is the tendency to be careless as opposed to deliberate and thoughtful.[73–75] Dickman[73] distinguishes two kinds of impulsivity: dysfunctional and functional. The former is the tendency to act with less forethought than most people, while the latter is the tendency to act with relatively little forethought when such a course of action is acceptable or optimal.

Prioritization (Machine[76]):The process of deconfliction or deferral of competing subsystem claims for cases of CPU, memory, and bandwidth overload. Given limits on system resources, rules or algorithms will be required in order to decide where resources will be allocated, and may be dependent on the system and mission.

Flexibility (Human and Machine): The ability to choose among approaches to tasks—or algorithms and processes, in the machine case—including multi-tasking and diverting—,for example, taking a different approach when the current one is not succeeding. In the machine case, as noted by Bryson,[77] the visible portions acting on the world are built-in action selection mechanisms in the agent architecture.

Adaptability (Human[78] and Machine): The ability to make appropriate responses to altered or changing situations; the ability to adjust one's behavior in response to different circumstances and/or different people. See Martin[79] for further discussion on conceptual bounds of this construct for human agents and Martin et al[80] for validation of an adaptability scale.

Intelligence (Human): The ability to derive information, learn from experience, adapt to changing environments, understand, and correctly use thought and reason.[78] See, also, the works by McGrew[81] and Sternberg[82,83] for psychometric approaches to intelligence. There is no generally agreed-upon definition of intelligence; it here refers to the many executive and related functions that may impact T&E.[84,85]

Computation Environment/Memory (Machine): The software, hardware, and firmware enabling computations, as well as any "controller" software, specific to autonomous operation or teaming. This can include a range of processes for accessing, interfacing with, and engaging with hardware components.

Physical Abilities (Human): The capability to perform some physical activity. Physical ability is defined by the requirements of a given task, occupation, or mission. One may possess physical ability to excel in performance in one occupation but lack the physical ability to perform in another. See Hogan[86] for further discussion.

Standard Platform (Machine): Platform hardware features make it possible for the machine to receive, process, and transmit information and interact with the world, including aspects required for integration. These features allow for computation, sensing, communications, and other interactions that include movement, engines, tires, wings, radar, and audiovisual sensing.

Physical Fitness (Human): Agent features related to strength, speed, and endurance.

Structural and Mechanical Elements (Machine): The machine agent's structural components such as frames, bearings, and axles; control mechanisms such as gear trains, brakes, and engines; control components, including actuators and controllers for other systems; and sensors that allow for integration. Fleming et al.[87] present a good discussion of these concepts for automotive applications.

Sensors (Human and Machine): Term for all input devices—sense organs and their capabilities in the case of human agents, all data intake devices not in Integration Hardware and Computer & Peripherals in the case of machine agents.

Computer & Peripherals (Machine): These are software and hardware components supporting additional mechanics not related to computation software. This can include hard drives and I/O hardware for platform operation. In some architectures, the teaming-specific function will be handled with common processors with platform operation, as discussed, for example, by Whitten & Bentley.[88]

Integration Hardware (Machine): Machine components that allow the subsystems on the platform to function together. This category overlaps with I/O hardware, wireless capabilities, and others needed for successful teaming.

Worldview (Human and Machine): The agent's resource-constrained processes to represent the current and changing state(s) of the environment(s), and the representation that results.

Judgment and Attitudes (Human): Judgment refers to the process by which multiple, fallible, and potentially conflicting cues are integrated to infer or deduce what is happening in the external world, as well as the inferences made by that process.[89] Attitudes are evaluations of beliefs that guide individual agents to adopt particular intentions, and are based on—but distinct from—the informational foundation of perceptions and beliefs.[50] Similarly, attitudes guide—but are not equivalent to—human intentions and, thus, behaviors.

Perspective (Machine): What the machine knows and/or believes about the world. This is consistent with the technical definitions from art or drawing, with explicit acknowledgment that a perspective can be misleading. According to Siciliano & Khatib,[90] this largely maps to the *sense* portion of robotics' sense-plan-act design philosophy that allows for planning and acting (i.e., decision making in our framework).

Situational Awareness (Human and Machine)[46,47]: The agent's perception of circumstantial and environmental elements and events with respect to time, space, and other agents (friendly, adversarial, and neutral). This includes the comprehension of the meanings of those elements and events and the projections of their future states. In the specific case of machine agents, there may or may not be an element of cyber situational awareness, as well, which can be defined as the collection and processing of data to provide an understanding of cyber health or compromise; see, for

instance, the definitions of Jajodia et al,[91] including the Identify, Protect, Detect, Respond, Recover framework.[92]

Trust (Human and Machine)[50]: The attitude held by an agent that another agent will, in good faith, help achieve goals in a situation potentially characterized by uncertainty and vulnerability. Lee and See's three general bases of trust[50] in the context of automation—performance, process, and purpose—are very likely to play a role in human-machine teams. Based on previous trust in automation research, levels of trust may be ill-calibrated (i.e., overly trusting or insufficiently trusting) due to erroneous judgments of a machine teammate's performance, process, or purpose.[51,93]

Cognitive Allocation (Human): A state in which cognitive resources are focused only or primarily on certain aspects of the environment and task at hand, and the system is in a state of readiness to respond to stimuli. This is adapted from the APA Dictionary of Psychology definition of *attention*. Because humans do not have an infinite attention capacity, the choice of metrics in this category will be determined by which factors most influence attention understanding in a given context.

Resource Allocation (Machine[94]): Processes that enable resource claims on machine operation (e.g., CPU usage, sensor operation), especially enabling teaming (e.g., explore-exploit tradeoffs in task completion and communication.

Working Memory (Human[95,96]): The cognitive system or group of cognitive systems that are used to keep necessary information in mind to enable the performance of complex tasks such as reasoning, comprehension, and learning. Many measures of working memory have been developed.[97,98]

Attention Allocation (Human[99]): The focused direction or distribution of limited cognitive resources. Humans do not possess an infinite capacity to attend to stimuli; attention allocation thus captures the direction or distribution of stimuli that cognitive resources are devoted to for a given period of time.

Process/Threat Monitoring (Machine): The I/O streams that feed normal operations and may be associated with cyber intrusions such as computer, user, access, SIEM, external, certificates, credentials, and other details for security assurance—,for example, Splunk, Snort, Wireshark, and so on.

Other Dependability Monitoring (Human and Machine): Processes—typically implicit, though not always—used for health monitoring systems and indicated mitigations. This includes monitoring of teammate, safety and reliability, legal/moral/ethical considerations, and assurances that warfighters and/or machines will not only do what you want but also not do what you do not want. In the machine case, this may include human health monitoring systems.

Effort (Human): The expending of physical or mental exertion by a human agent.

Resource Use (Machine[90]): Resource use in machine operation (electrical power, CPU usage, sensors, integration of inputs). Measurements required for the machine to function would include the following measures, most of which are lower-level processes that enable the dependable functioning of complex mechanical systems.

Resource Availability (Human): Umbrella category for workload, fatigue, and resource accounting.

Platform Operating Margin (Machine): Size, weight, power, and cost (SWaP-C) limits; analogous to limits on non-autonomous systems.

Workload (Human[100]): The portion of operator physical or mental resources that are required to meet task demands—,ithat is, the costs to the operator to accomplish the mission. Workload might include the amount of work such as the number of tasks to complete and the difficulty of those tasks, the time that one has to complete the task, and the subjective experience of the human operator.[100,101] Thus workload might include objective and subjective measures, but it is often measured using a subjective scale of workload that measures the operator experience of workload, such as the NASA-TLX.[102,103]

Fatigue (Human[104]): Fatigue is a form of inadequacy in which the agent experiences an aversion to exertion and a sensation of an inability to continue activity. It may be physical, cognitive, or affective (i.e., manifested by low mood or lethargy).

Processing (Machine): CPU and memory margins, as well as thermal and power constraints on processing, the latter of which could also apply to Platform Operating Margin above. Note that the monitoring in "resource allocation" is a "demand" signal on resource use for the sensors and other subsystems. Resource use means the provided resources, whatever the demand.

Usability (Human-to-machine): This is one of the only class of metrics that, in our assessment, had a clear human agent connection without a machine analog, primarily because usability refers specifically to a human's perception of the ease with which an operator can use a machine system (intelligent or otherwise). There is a great deal of literature devoted to usability, and it is a complex, multi-dimensional concept. Usability criteria will vary from system to system. Nielsen[105] provides a broad starting point for usability discussions. The System Usability Scale (SUS) is generalizable to a wide range of systems and is often used as a subjective measure of usability.[106]

Performance (Human and Machine): Assessment of decisions, results, and subsequent effects generated by or attributable to agent action.

Decision Making (Human and Machine)[89]: The process of choosing or determining a course of action, including the consideration of alternative courses of action, uncertain circumstances and conditioning events, and consequences associated with potential outcomes. In the machine case, Veres et al.[107] discuss various autonomous forms of action and decision making, as well.

Optimality (Human and Machine): Defined as at the team-level, with the scope narrowed to a single agent.

Robustness (Human and Machine)[55]: Defined as at the team-level, with the scope narrowed to a single agent.

Risk Level (Human[108]): The extent to which a decision involves a gamble on its desired outcome or expected value as opposed to ensuring that desired or expected outcome. Metrics of risk for human agents must take into account the fact that humans,

as opposed to machine agents, are notoriously bad at intuitively assessing risk and probabilities, but are also generally risk averse in their decision making,[89] making for some contradictory situations. For instance, as Kahneman & Taversky[108] have noted, when it comes to negative outcomes or losses, humans tend to prefer maximizing the probability of avoiding loss even if the expected value of those decisions are lower (e.g., preferring have a 50% chance of losing $100 and 50% chance of losing nothing over a certain loss of $45).

Risk Level (Machine): Machine agents' assessments of risk, danger, and so on. are very likely to be different from humans for the foreseeable future, so evaluations of agents' determination of risk should take that difference into account. Indeed, design of autonomous machines can attempt to bias the machines' decision making in either risk-tolerant or risk-averse fashions, leading calibration to be important for T&E. Incomplete understanding of contributors to the decision making by designers, operators, commanders, or teammates may result in a risk tolerance incompatible with team's intent. This is a subject for T&E determination.

Reliance (Human and Machine)[50,109]: The behavioral act of dependence on another agent for the accomplishment of some goal. Reliance is influenced by trust but not determined by it.

Task Performance (Human and Machine)[110]: The effectiveness with which agents engage in and complete activities that contribute to the overall mission either directly, such as through implementing an action or decision, or indirectly, such as by providing necessary information or services. Task performance should be distinguished from contextual performance, which includes factors like trust and influences effectiveness in more abstract and subjective ways that determine and characterize the relevant organizational, social, and psychological contexts. Task performance is best measured using objective factors such as those described below.

Error Rates: The frequency or duration of deviations from correct task procedure throughout mission accomplishment relative to total tasks or task time performed for the mission. Examples of error rate measures include the number of errors in a given time period, and the probability of a miss given the number of shots taken.

Timeliness (Human and Machine)[111]: The extent to which to which the agent completes the required task(s) within an allotted amount of time. Examples of timeliness measures include time to complete a task, time to target, and time to detect.

Persistence (Machine)[112]: The action of persevering, or the sustainment of directed effort on a task or toward a goal despite obstacles. Persistence has a positive connotation and persisting behavior has been shown to enhance performance.[113] Despite this, in some instances this behavior can impede successful completion of a higher goal.[114]