The Design of an Ostensible Human Teammate

Elizabeth L. Fox*
Air Force Research Laboratory
Wright-Patterson AFB, Ohio
elizabeth.fox.9@us.af.mil*

Gregory Bowers
Ball Aerospace
Dayton, Ohio

August Capiola
Air Force Research Laboratory
Wright-Patterson AFB, Ohio

Arielle L. Stephenson
Ball Aerospace
Dayton, Ohio

Abstract

Reliance on computer-mediated teaming has exploded in recent years, making research on how teammates calibrate their behavior critical. Here, we offer a simplistic, viable method to model human behavior for use in subsequent research investigating coordination among partners. We collected human performance data in a multiple object tracking task and a communications task to serve as the basis of our agent performance in multiple tasks. We demonstrate our model in real-time by drawing from existing research involving probabilistic models of detecting critical events and sample from a parametric log normal model of human response times to mimic human behavior. We endow our agent with team-based etiquette through a hesitancy to intervene, a parameter sampled from a uniform distribution, and manipulated agent performance through parametric shifts to detection and the log normal distribution that represents agent response times. The present work does not offer hypotheses as we did not conduct an experiment. Rather, we derive and provide a validation of an agent modeled from human performance parameters in two tasks for future team-level research with ad hoc partners.

Keywords: ostensible teammate, agent, human performance, parametric models

1. Introduction

Advancements in technology have restructured the way individuals understand and perform tasks. In the 21st century, most work domains heavily rely on computer-mediated tasks – i.e., workers accomplishing goals through a computer or technological system, rather than physically interacting with the environment. This presents new possibilities for workers and teams utilizing such methods. For example, it enables one's interaction with remote environments and geographically separated teammates to accomplish tasks. However, for industry, government, and academia to fully leverage these opportunities, research must be done on realistic coordination in distributed contexts. Controlled team-based research is difficult; it requires different (but consistent within a range) behavior and performance. That is, experiments manipulating features of human behavior and performance are difficult to produce. However, one workaround is to create ad hoc agents that can replicate human behavior. This allows researchers to conduct team-level experiments without recruiting and coordinating multiple participants at once.

The present work prioritizes this with a first step—developing a realistic agent with performance parameters for use in research on computer-mediated coordination amongst ad hoc partners. In the following pages, we provide a brief summary of the need for effective distributed teaming and provide an example on the effects of this kind of research on human psychology and performance. Then, we outline the development of an agent for use in future research on distributed teaming and leverage human performance data instantiating task-relevant parameters, their variability, and their effects on ad hoc coordination. Finally, we discuss future research ideas.

1.1. Realistic Research on Distributed Teams

Researchers, managers, and business owners must consider the benefits and costs of integrating existing and emerging technologies to facilitate effective distributed work. Compared to traditional organizations comprising face-to-face interactions between co-located parties, organizations have evolved to comprise multiple dispersed teams (Benishek and Lazzara, 2019). Geographically separated, or distributed, teams are advantageous to occupy multiple operation locations and domains, e.g., Joint-All-Domain Command and Control (JADC2) operations within the United States Department of Defense (DoD; Congressional Research Service, 2020) prevent operations from containing a single point of failure (Priebe et al., 2020). With the emergence of geographically distributed collaborations utilizing technology as means of communication

URI: https://hdl.handle.net/10125/102657
978-0-9981331-6-4

(Morrison-Smith and Ruiz, 2020), the advantages of distributed teams are prominent in that the location of specialists is no longer a limitation.

Many of the underlying dynamics of human-human teams can carry over from co-located to distributed teams (Corbitt et al., 2004), but, importantly, the nature of the team context shifts in distributed contexts; for example, distributed and/or ad hoc partners often do not have a history of working together, yet still must work together in order to achieve their common goal(s). In such instances, people may seek out anecdotal information (e.g., "Your partner has a good (bad) reputation") to form judgments about one another prior to working together. These judgments may provide partners the means to calibrate subsequent behavior, and ultimately influence team performance (Capiola et al., 2020). Carefully choosing what and how to provide information in distributed team contexts is important: it can impact psychological constructs such as trust (e.g., Capiola et al., 2020; Kanawattanachai and Yoo, 2002) or collective efficacy (Capiola et al., 2019), factors which facilitate performance in distributed teams.

More research and development is necessary to understand the implications of distributed teaming on performance, how to train for effective collaboration in human-human partnerships in distributed domains, and how information is most appropriately relayed between parties to maximize efficiency in distributed contexts. The present work prioritizes building a simplistic, but plausible, agent to stand-in as a teammate such that researchers can investigate factors (e.g., pieces of information about a partner) which influence psychological constructs, behavior, and performance in distributed, ad hoc teams.

1.2. AI as a Human Stand-in

Practically, the coordination of human-human team research is laborious in both co-located and distributed contexts (Mathieu et al., 2008). Artificial teammates, or agents, are presumed to be useful for assisting individuals and teams, once their functionality is optimized and their integration is accepted by humans (Chen and Barnes, 2014). In laboratory experiments, researchers have simulated intelligently designed agents by pre-programming actions participants view as being attributed to an agent (Alarcon et al., 2022) or employing wizard of Oz techniques (McNeese et al., 2019) to serve as stand-ins for a human teammate to investigate individual (e.g., risk-taking behaviors) and team-level constructs (e.g., coordination), respectively. The reverse has seen some study, evaluating the capability of agents to pass as human and the results of betrayed expectations in performance (Grimes et al., 2021). Human-like responses and interaction encourage people to engage with the agent as they would a human. This is important as it may provide the realism and fidelity necessary for researchers to capture human-human team dynamics through a model-based agent.

Research with the development and use of cognitive models shows agents can successfully reproduce human behavior and effectively mimic human-human interactions in complex and real-world tasks (Rickel and Johnson, 1999). While real-world distributed team contexts may include live interactions that could prove difficult to artificially reproduce (e.g., face-to-face video, verbal communication), many tasks do not – or cannot – explicitly require those features. Hence, intelligently designed agents are a plausible and worthwhile endeavor for research and development of team coordination and performance in distributed teams. Further, specific types of human performers may be difficult to both recruit and coordinate their participation in human-subjects experiments. This makes it extremely difficult to investigate important research questions and quantify effects of exogenous variables onto endogenous variables.

Human-human and human-machine interactions differ; therefore, it is important to create models in which human behavior serves as the basis for building intelligent synthetic agents. Cognitive modeling frameworks such as SOAR (Laird et al., 1987) and ACT-R (Anderson et al., 1997) can be used to develop tools such as tutoring software (Ritter et al., 2007) and performance support systems (Lovett et al., 2000). However, in the present work, we use a joint model of speed and accuracy to mimic human performance and make a few assumptions about how teammates interact with one another to reduce the overall development time and computational demands of creating and applying the intelligent agent in real-time. Our model needs to mimic human behavior, performance, and team interaction to effectively stand-in as a teammate in future research on human-human teams.

Next, we define agent-based performance that can generalize to many tasks, with various attributes such as assertiveness or accuracy derived from parametric models of human performance. Then, we describe two tasks, a Multiple Object Tracking (MOT) and a Communications (Commms) task, to demonstrate our agent through real-time simulation and parameter recovery, which may be used in future research.

1.3. Agent Model

Agent performance can be based on a parametric model of human performance. For a high performing agent, we can assume response times (RT) follow a log normal distribution, $lognorm(\mu, \sigma^2)$, of human response times, x_h, such that the probability density function (PDF) of agent RTs, $f(x_a)$, is defined as:

$$f(x_a) = \frac{1}{\sigma_{x_h} x_h \sqrt{2\pi}} \exp^{-} \frac{(ln(x_h) - \mu_{x_h})^2}{2\sigma_{x_h}^2},$$

where x_h are observed human RTs and μ_{x_h} and σ_{x_h} are the expected mean and standard deviation of the natural logarithm of the human data, x_h, respectively. In this paper, we referred to each parameter as $log(\mu)$ and $log(\sigma)$.

We use the maximum likelihood estimate (MLE) to select parameters at the group-level, $log(\mu, \sigma)$ for each task type and high or low performance, respectively. Monte Carlo simulation and a MLE fit on shifted response times served as the model of detection time for a low performing teammate, $f(y_a)$, where low performance *shifted* data, y_s, is characterized as the human data (representing a high performing agent), x_h, plus two standard deviations of the human data,

$$y_s = x_h + 2 * \sqrt{\frac{\sum_{i=1}^{n} (x_{h_i} - \bar{x_h})^2}{n-1}}$$

where n is the number of samples of response times and $\bar{x_h}$ is the average of the human response times. Similarly, the PDF of low performance agent RTs, $f(y_a)$, is defined as:

$$f(y_a) = \frac{1}{\sigma_{y_s} y_s \sqrt{2\pi}} \exp^{-} \frac{(ln(y_s) - \mu_{y_s})^2}{2\sigma_{y_s}^2},$$

where μ_{y_s} and σ_{y_s} are the expected mean and standard deviation of the natural logarithm of the shifted, or 'low', human performance data, y_s, respectively.

Existing literature informed the detection rate (i.e., accuracy: hit/miss) of agents that represented either high or low performing teammates (e.g., Dixon et al., 2006, 2007; Parasuraman and Manzey, 2010; Rice and McCarley, 2011). We refer to these detection rates as d_x and d_y for high and low performing agents, respectively. In this model, if the agent can assist, i.e., detects human should respond, and an adequate amount of time passes, then the agent will respond for the participant. In this simple model, the agent never makes an incorrect response; that is, mistakes are only expressed as missed events. Therefore, the PDF for RTs of, for example, a high performing agent conditioned on its probability of detection, $f(x_a|d_x)$, is characterized as:

$$f(x_a|d_x) = \begin{cases} f(x_h) \text{ for } d \leq d_x, \\ \text{NA for } d > d_x \end{cases}$$

where *NA* indicates a failure to detect a response is warranted, recorded as a *miss*; d is a random integer between 0-100, sampled at the start of each trial; d_x is the detection rate for the high performing agent. Similarly, the PDF for RTs of a low performing agent conditioned on its probability of detecting an event, $f(y_a|d_y)$ is characterized in the same way, where $f(y_a)$ and d_y represent the RT distribution and detection rate of the low performing agent, respectively. Additionally, we implement a lag time, l, to intervene using a continuous uniform distribution, $U(m, n)$, with two parameters, m and n, which represent the minimum and maximum bounds on the lag time value, respectively. As such, the PDF of lag time is defined as:

$$f(l) = \begin{cases} \frac{1}{n-m} \text{ for } m \leq l \leq n, \\ 0 \text{ for } l < m \text{ or } l > n. \end{cases}$$

Therefore, the intervention time, IT, the time in which the agent *may* intervene, is the sum of the detection time and lag time, $f(l)$. For example, the PDF of intervention times for a high performing agent is defined as:

$$f(IT)_{x_a} = f(x_a) + f(l)$$

where $f(x_a)$ represents the $lognorm(\mu, \sigma^2)$ of the high performing agent. Similarly, intervention time of the low performing agent, $f(IT)_{y_a}$, is captured by the sum of the low performance PDF, $f(y_a)$, and function of lag time, $f(l)$.

Importantly, the agent only intervenes should it detect a response is desirable, where detection rates differ depending on whether the agent is a high or low performer, AND (denoted as $\wedge$) the human did not yet respond at the time of trial onset plus the intervention time, IT. Logically, it follows that the agent does not intervene should it not detect a response is warranted OR (denoted as $\vee$) the human responded before the sampled intervention time. Therefore, the intervention time on any given trial, IT_i of, for example, a high performing agent is:

$$IT_{x_a i} = \begin{cases} IT_{x_a i} \text{ if } d_i \leq d_x \wedge IT_{x_a i} < h_i, \\ \text{NA if } d_i > d_x \vee IT_{x_a i} \geq h_i \end{cases}$$

where *NA* indicates no response from the agent; d_i is the random integer between 0-100 generated on trial i; and h_i is the response time of the human partner (if a response was made) on trial i. Similarly, observations of IT for a low performing agent conditioned on its detection rate is characterized in the same manner, where y_{ai} and d_y represent a response time and the detection rate of the agent, respectively.

To the authors' knowledge, existing literature does not provide direct suggestion regarding the extent to which humans wait before intervening in their partner's task, though tangential work does exist. For instance, team delay with respect to toleration of connectivity lag between human teammates in a video game setting provides minimal discussion regarding levels of peer interlude (Saint John and Levine, 2005). Here, team play is examined by leveraging a radius, dubbed team radius ($t = l + b$), with a sphere of influence (SOI) being $l = 1$ and a boot radius ($b = 2$), resulting in a delay tolerance of about 25ms. Further, research investigated the malleability of human collaboration with a robot when asked to move a table through a doorway and presented with a strategy that came from the robot (Nikolaidis et al., 2017). Here, adaptability was defined as the probability the human would move from their solution to the provided one, accounting for individual differences amongst each person. When jointly performing the task with a human partner, the robot offered an option to achieve the task in increments of one second, starting immediately. More willing human partners formed an agreement to use the robotic solution at one second while a majority of human partners accepted the robotic solution at three or more seconds, regardless of identified adaptability levels.

Despite the lack of guidance from extant literature, we choose m and n parameters that provide adequate delay in order to effectively capture human hesitancy and nature of maximum payout (i.e., it is better for each teammate to perform their own tasks, but intervening is better than a miss/incorrect response). In the current design, samples of lag time were generated from the same distribution for both high and low performing agents.

2. Methods

In order to create human-like responses we chose two tasks to demonstrate our agent: a Multiple Object Tracking (MOT) and a Communications (Comms) tasks. We manipulated the probability the agent will detect an event occurred (i.e., hit rate) and, if detected, the speed at which the agent responded. The speed and correctness of the agent varied, depending on whether it emulated a high or low performing person. In the next section, we outline and sketch a generalizable MOT and Comms task. Then, we report both our method to estimate parameters that represented agent performance and the procedure we used to collect human performance data in these tasks.

2.1. Summary of Tasks and Application

In the present instantiation of our agent, we endow it with unique responsibilities in two subtasks, the same two tasks its human partner simultaneously completes. Should it both detect assistance is desirable in either task (i.e., detect an action is necessary), and that its human teammate has not completed said task(s) within a comfortable amount of time, it occasionally assists its human teammate. The first task is a MOT task where the agent is responsible for turning on/off an alarm in one high-risk area of interest (AoI). This AoI is one of four quadrants of a display (the diagonal quadrant is the human partner's AoI and the other two are low-risk cells). The correct state of the alarm depends on the ratio of hostile (red) versus security (green) actors in the designated AoI (blue dots represent benign bystanders). On the same shared display, the human partner will complete the same MOT task, but will do so for a different AoI (see Figure 1). The other task is a Comms

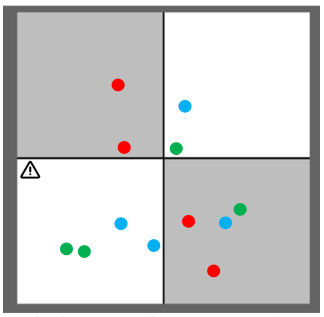

Figure 1: A static image of a MOT task. Color and size of tracks are changed and enlarged, respectively.

task where the agent "listens" for their assigned call-sign (e.g., agent = "Alpha") and responds to the message that follows their call-sign. A message is a string of 4 spoken letters, for which the agent must respond with the letter that follows the 3rd letter in the string in the English alphabet. For instance, in a scenario where the agent is responding to the message "B, D, R, M", the correct response is the letter "S", as it is the next letter of the alphabet following the third letter in the spoken string (see Figure 2). The human partner will complete the same Comms task, but will respond to a different call-sign (e.g., human = "Bravo"). Distractor call-signs (e.g., "Delta") were also present. The basis

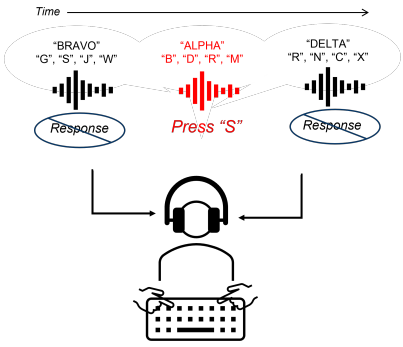

Figure 2: A representation of the Comms task.

of our agent model is formed using the combination of performance of several human participants that are well-practiced in the task, assumptions about what constitutes 'high' (fast and accurate) or 'low' (slow and inaccurate) performance, and social etiquette norms of teammate intervention and assisting behaviors in distributed and multi-task team contexts. Next, we unpack the instantiation of these parameters for our use case.

2.2. Performance

We manipulated both detection rate and speed for our high and low performing agents. Our detection rate parameters were determined by research examining the effects of high/low performing agents in human-machine teaming literature (e.g., Dixon et al., 2006, 2007; Parasuraman and Manzey, 2010; Rice and McCarley, 2011). Specifically, the probability that our agent detected an event (hit rate) was low (65%) or high (95%). When an event was triggered, a random sample from 0-1 was chosen – if the random number was lower than the probability of hit (i.e., $x < 0.65$ for low, $x < 0.95$ for high) then the agent attempted to respond to that event; if the random number was higher than the probability of the hit (i.e., $x > 0.65$ for low, $x > 0.95$ for high) then the agent would fail to detect that an event occurred. If the agent did respond, they were always correct (i.e., no false alarms). If the agent attempted to respond to an event, a detection time was randomly sampled from our model based on human performance in each isolated (MOT and Comms) task. We explain the human data and response time (RT) model in the following sections.

Our current instantiation of the model does not include false alarms; therefore, incorrect responses were only observed through the absence of a response, or miss. However, this is a modeling choice for our specific example described in this paper. Modelers interested in creating an agent that produces false alarms can easily include a parameter to represent the probability in which the agent may detect an event at any given time, or in response to events that may occur but do not require a response (i.e., distracting events).

If the agent detected an event, we on-the-fly took samples from a model that was fit to data we collected from human subjects ($N = 5$; note that, in our case, this is a sufficient sample size for obtaining parameter estimates due to low variability among subjects and large number of observations per individual), three and two of whom were coauthors of this paper and members of the lab, respectively. The model fit from human data represented the distribution of potential RTs for the high performing agents. We added two standard deviations to the human data and fit a different model to these data to represent the distribution of RTs for the low performing agents.

2.3. Human Subjects Data

Human subjects data collection consisted of people completing three blocks of only the Comms task and two blocks of completing only the MOT task. The tasks were completed at the lowest level of difficulty such that each human subject was only responsible for one call-sign or one quadrant. Hence, our agent performance was modeled as a reflection of ideal human circumstances, an isolated task context in a distraction-free environment. Experiment blocks were kept short as to not introduce effects from fatigue or sustained attention. All human subjects were well-versed in the task and sufficiently trained prior to completing the experimental blocks. Each human subject was assigned each call-sign once and each quadrant once, alternating between Comms and MOT blocks. For example, a human subject may complete blocks in a pattern of Comms (Alpha), MOT (top left), Comms (Bravo), MOT (bottom right), and Comms (Delta). Each Comms block consisted of 50 trials (roughly 6-7 minutes) and visual blocks were 6 minutes. This provided 50 trials per subject of 'hit' trials for the Comms task and around 160 events per subject in the MOT task.

2.4. Model Fit

2.4.1. Multiple object tracking (MOT) task We were conservative in extracting MOT task data to use for modeling our simulated agent performance. Therefore, we excluded any instances in which a human subject turned ON and OFF an alarm within a single *UNSAFE* period and any instances in which the total duration of

an *UNSAFE* state was less than one second. We only considered RTs that were correct and collapsed across all human participants to fit a single model to the group data.

High performance. We fit a log normal model to the RT data for the MOT task. Figure 3a shows the individuals' data for the MOT task (colored lines): the black line is the group data, and the dashed line shows the model fit using a Monte Carlo sampling method. The legend shows how many data points were used to estimate each human participant's probability distribution.

The log normal parameters for our high performing agents in the MOT task were $log(\mu) = -0.727, log(\sigma) = 0.370$. There is some chance that a negative RT is sampled. In this case, the model immediately resamples from this distribution until a positive RT is obtained.

Low performance. We added two standard deviations of the group data to each raw data point in order to create our low performing model. This shifted the response time distribution to generally be slower than the raw data. We fit a log normal to these data; the log normal parameters for our low performing agent in the MOT task were $log(\mu) = 0.853, log(\sigma) = 0.255$. Figure 3b shows samples from the high (green line) and low (red line) performing model.

In future experiments, the response period that a participant or agent will have to respond to turn ON/OFF an alarm may be random and depend on the position of the dots. Therefore, using the human data, we calculated the duration that the AOI in the MOT task was in a particular state and compared it to the simulated agent performance. This was to check that the agent would indeed turn ON/OFF the alarm within a reasonable time in both the high and low performance conditions. The black line in Figure 3b shows the response periods relative to high (green) and low (red) performing agent responses.

2.4.2. Auditory communications (Comms) task Our approach to the Comms task data was more straightforward. We did not exclude any correct response times and collapsed across assigned call-signs for each human participant.

High performance. Similar to the MOT task, we fit a single model to the group data and obtained two parameters for the log normal distribution ($log(\mu)$, $log(\sigma)$). We used Monte Carlo sampling ($N = 1000$) to illustrate the model fit, shown in Figure 3c. Again, each individual's data are indicated by colored lines, the group data is the solid black line, and the simulated agent data is the dashed black line. The legend shows how many data points were collected for each participant, collapsed across each assigned call-sign (Alpha, Bravo, Delta).

The log normal parameters for our high performing agent in the Comms task were $log(\mu) = 0.303, log(\sigma) = 0.361$. Similar to the MOT task, if the RT sampled from the distribution is negative then the model immediately resamples from this distribution until a positive RT is obtained.

Low performance. We added two standard deviations of the group data to each raw data point in order to create our low performing model. This shifted the RT distribution to generally be slower than the high performance data. We fit a log normal distribution to these data; the log normal parameters for our high performing agent in the Comms task were $log(\mu) = 0.972, log(\sigma) = 0.202$. Figure 3d shows samples from the high (green line) and low (red line) performing model. The cutoff period for responses was fixed (as opposed to variable in the MOT task) to 3 seconds, shown as a gray vertical bar.

2.5. Intervention

We created an agent that will detect state changes (in the MOT task) and pertinent call-signs (in the Comms task) in the participant's tasks with the same probability that it detects events occurring in its own tasks. For example, the participant may be assigned to monitor the top left AOI (MOT task) and "Bravo" (Comms task), and the agent may be assigned the bottom right AOI and "Alpha." A low performing agent detects 65% of the MOT task state changes (i.e., turn ON/OFF an alarm) or Comms task responses (i.e., participant's call-sign was spoken) in either AOI (top left, bottom right) or call-sign ("Alpha", "Bravo"). A high performing agent detects 95% of the pertinent AOIs or call-signs in their own and the participant's assigned MOT tasks and Comms tasks. Similar to an agent's response to events in which it was responsible for monitoring (bottom right AOI and "Alpha"), if an event was detected in the participant's tasks (top left AOI and "Bravo"), then a response time is generated by sampling from the appropriate log normal distribution given the agent's set performance level (low, high) and the task (MOT, Comms).

Next, we incorporated an intervention delay parameter for the agent, ID. In practice, if the intervention delay time passes and the human teammate still does not make a response, then the agent will respond for them. The ID is task dependent: the total response period for the MOT task will vary depending on how long the dots remain in the AOI, as evidenced by the black distribution in Figure 3c; the response period

for the Comms task is consistently 3 seconds after the onset of the 3rd letter, as shown by the gray vertical line in Figure 3d. Given these constraints, the lag time for the MOT task was sampled from a uniform distribution ranging from $1-2$ seconds; the lag time for the Comms task was sampled from a uniform distribution ranging from $0.5-1$ second. The response time for the agent to respond to events in the participant's task depends on the overall agent's performance level and the task; simulated intervention RTs are shown in Figure 3e and 3f.

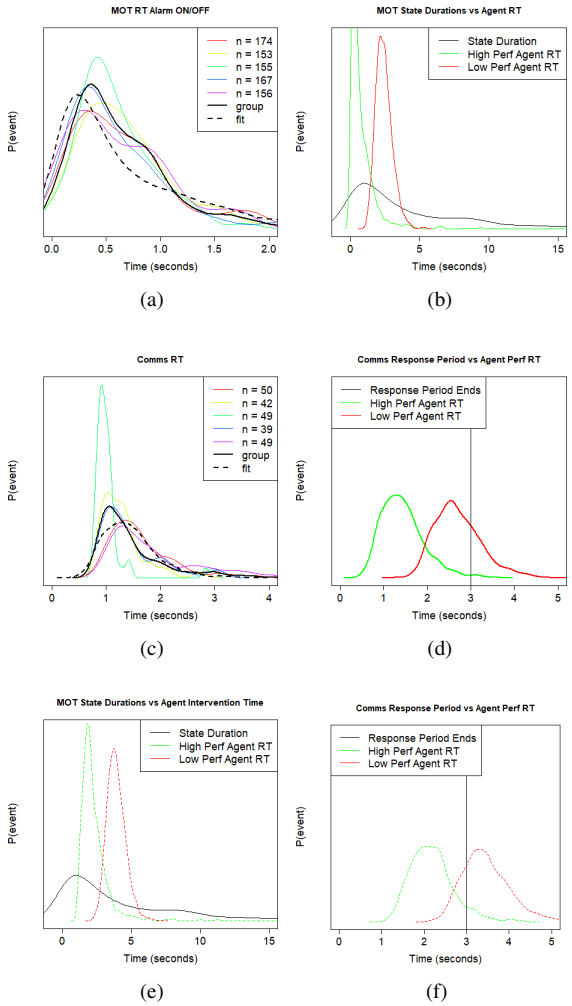

Figure 3: Individuals' data for the MOT (a) and Comms (c) task, group data, and model fit. The legend indicates the number of data points per distribution. Trial duration relative to high and low performing agent in the MOT (b) and Comms task (d). Simulated intervention RTs for a high and low performing agent in the MOT (e) and Comms (f) task, relative to the response period.

3. Model Validation

We validated the performance of our agent by recording its responses in a modified version of the program which ran continuously (generating new trials) until 1000 responses in each category (visual + agent, visual + human, auditory + agent, and auditory + human) were complete. One data object stored all attempts (any time a valid call sign or quadrant occurred, and thus was checked against the agent to see if it would respond). Another stored the "delay" times generated for the agent when it actually did respond.

The total number of responses (for each category) was divided by the total number of attempts to determine the real probability, which was compared to the aforementioned values. Additionally, the delay times were evaluated by determining the μ and σ values (mean and standard deviation of the natural logarithm of the delay times). The results for this are in Table 1 and 2.

As can be seen, all observed values are very close to the defined parameters. The categories which involve the intervention delay (visual + human, auditory + human) show average lag values, in each case, very close to the middle of the uniform distribution. As with any probabilistic system, more samples would asymptotically improve the accuracy of the results, but as can be seen here 1000 observations are sufficient to show our agent teammate performs as anticipated.

4. Discussion

Our goal was to develop a cognitive-based, real-time agent that is a viable distributed, ad hoc partner for future work investigating team-level psychological constructs and performance in laboratory settings. We detailed how we created and demonstrated our agent for two tasks, a MOT and a verbal Comms task. We successfully validated our model by simulating performance in each task and recovering the response time and accuracy parameters generated from human subjects. This simple, yet viable, human performance-based agent can stand-in as a human teammate for future laboratory studies investigating computer-mediated collaborations. Our methodological approach can be adopted by researchers wishing to develop their own agent(s) for team-based work.

One strength of our modeling technique is its simplicity. We collected data from five human subjects, and in a subset of conditions, to fit a straightforward model of response times, containing only two parameters in each task, and two more for when it assists with a human teammate's task(s). We used the lower bound of variation in our

Table 1: "Low" performance agent parameters

Expected/Defined	**MOT**	**Comms**
Response Probability (Self)	0.65	0.65
Response Probability (Intervention)	1.0	1.0
Response Delay (seconds)	$\log(\mu) = 0.853, log(\sigma) = 0.255$	$\log(\mu) = 0.972, log(\sigma) = 0.202$
Intervention Delay (seconds)	min.=1.0, max.=2.0	min.=0.5, max.=1.0
Observed	**MOT**	**Comms**
Response Probability (Self)	0.65	0.65
Response Probability (Intervention)	1.0	1.0
Response Delay (seconds)	$\log(\mu) = 0.856, log(\sigma) = 0.257$	$\log(\mu) = 0.970, log(\sigma) = 0.203$
Intervention Delay (seconds)	$\mu = 1.50$	$\mu = 0.75$

Table 2: "High" performance agent parameters

Expected/Defined	**MOT**	**Comms**
Response Probability (Self)	0.95	0.95
Response Probability (Intervention)	1.0	1.0
Response Delay (seconds)	$\log(\mu) = -0.727, log(\sigma) = 1.123$	$\log(\mu) = 0.301, log(\sigma) = 0.361$
Intervention Delay (seconds)	min.=1.0, max.=2.0	min.=0.5, max.=1.0
Observed	**MOT**	**Comms**
Response Probability (Self)	0.95	0.95
Response Probability (Intervention)	1.0	1.0
Response Delay (seconds)	$\log(\mu) = -0.709, log(\sigma) = 1.125$	$\log(\mu) = 0.336, log(\sigma) = 0.365$
Intervention Delay (seconds)	$\mu = 1.50$	$\mu = 0.75$

human performance data to simulate slow, or 'bad', performance. We drew from existing literature where possible to estimate detection rates of a high or low performing agent and whether our agent would respond. We sampled from our parametric model of agent performance in real-time to determine the speed and accuracy of the agent in response to events that either the agent or human would be responsible for completing in the MOT and Comms task.

4.1. Limitations and Future Research

We purposefully did not instantiate agent false alarms when responding to their own or their partner's alarm or call-sign, i.e., the agent never responded when they should not have. Obviously, humans may offer such responses in a team-based task as described here. However, we did not qualify the quality of the performance manipulations (e.g., "My partner is not helpful because a) they miss a lot of suspicious personnel in my quadrant and/or b) they respond there is a suspicious person in my quadrant when there is not"). Research on human-machine teaming has documented the effects of misses and false alarms on human reliance and compliance with automation (Dixon et al., 2006, 2007). Future work may wish to model agent false alarm rate and qualify the kind of false alarm made to investigate its effect on psychological constructs and performance in human-(ostensible) human teams.

Additionally, our agent was modeled after specific values for how parameters of speed and accuracy increased or decreased depending on whether we wanted the agent to exhibit high or low performance. Good team performance depends on whether the human teammate can detect and calibrate their performance appropriately. Ideally, humans will utilize spare attentional resources to monitor and assist their teammate when necessary to achieve a level of performance that exceeds that of each alone. Nonetheless, how one obtains knowledge about their partner's performance may influence their interactions. For instance, forming a representation over time, 'learned knowledge', versus acquiring information prior to an interaction, 'explicit knowledge', may change behavior (e.g., Zhang and Houpt, 2020). Similar phenomena may generalize to human- (ostensible) human interactions.

Until now, no literature suggests how to model the hesitancy, or lag time, when intervening in a partner's task. We used a task assessment, e.g., average duration of response period, to determine an upper and lower bound to represent intervention time in each task. In order to calibrate our intervention time parameters, we had to first create an agent for people to work alongside and have the opportunity to intervene. Hence, our intervention lag parameter could be improved with additional research. Nonetheless,

our agent anecdotally passed the test of acting 'human' during an active demonstration in a preliminary lab study. Our live demonstration was done with willing and naive individuals under the guise of completing a task with separate human partners. In multiple independent instances, these individuals made comments related to the believably of our agent. This included declaring the desire to "not want to let down my partner" as well as exasperation when paired with a low performing agent. These individuals were informed after the demonstration that they were partnered with an agent for the tasks, after which the agent received praise for its convincing nature.

Our agent development consisted of performing a MOT and a Comms task. However, this does not limit applying our modeling technique to develop intelligent agent teammates in other types of distributed, ad hoc contexts. Similarly, our agent was split broadly into "high" and "low" performing variants, but more granular distinctions could be made for individual parameters. One might consider the parameters indicative of attributes like alertness (the probability of perceiving an event has occurred), accuracy (always 1 in our scenario, but still able to be manipulated), and assertiveness (the agent's likelihood of or lag time before intervening in the human partner's task). Such distinctions would allow for agent variants targeting different aspects of a team task. For instance, the agent performance in each of our tasks was modeled after the human subjects' *best* performance, meaning when they completed the task in a distraction-free and single-task context. Future work could extend our agent model to exhibit dual-task deficit that is expected when time-sharing between two challenging, or 'resource-limited', tasks (Norman and Bobrow, 1975). The MOT and Comms tasks were designed such that, according to principles of multiple resource theory (C. Wickens, 1984), the tasks have little to no overlapping resource demands, which minimizes their degree of competition for attention and predicted dual-task performance deficit (C. D. Wickens, 2002). However, when teammates attempt to complete multiple demanding tasks simultaneously (e.g., Fox et al., 2021) or complete their own tasks while monitoring, and sometimes intervening in, their partner's task, resource demands increase and performance deficits are expected. Future work should assess this.

4.2. Conclusions

We developed the theory, mathematical instantiation, and software of a parametric model to serve as the basis of an agent teammate. We demonstrated our model using data from well-practiced human subjects and parameter estimates provided in existing literature on automation detection rates. We imposed a hesitancy, or lag, when the agent intervened in their partner's task to mimic social etiquette when working in team-based contexts. We demonstrated our agent as a viable stand-in for a human in distributed and ad hoc contexts through model simulation and validation of parameter recovery. The simplistic nature of our model provides the opportunity for human participants to interact with the agent in real-time, allowing human-human teaming research to be conducted with only a single subject and an easily controlled partner. We highlighted gaps in current literature where our model could add an invaluable contribution and posed a few model improvements or ways that researchers could adapt our framework to accommodate various physical demands (e.g., different tasks), abilities (e.g., time-sharing efficiency), and important psychological constructs (e.g., trust) in future research.

4.3. Acknowledgements

The views expressed are those of the authors and do not necessarily reflect the official policy or position of the Department of the Air Force, the Department of Defense, or the U.S. government. No potential conflict of interest was reported by the authors. The research was supported, in part, by the 711 Human Performance Wing Chief Scientist Office (contract FA8650-20-D-6203). The study was approved by the Air Force Research Laboratory 711 Human Performance Wing Institutional Review Board (protocol FWR20220029E, V1.01). Distribution A. Approved for public release; distribution unlimited. AFRL-2022-2275; Cleared 12 May 2022.

References

Alarcon, G., Capiola, A., Morgan, J., Hamdan, I. A., & Lee, M. (2022). Trust violations in human-human and human-robot interactions: The influence of ability, benevolence and integrity violations. *Proceedings of the 55th Annual Hawaii International Conference on System Sciences*.

Anderson, J., Matessa, M., & Lebiere, C. (1997). ACT-R: A theory of higher level cognition and its relation to visual attention. *Human-Computer Interaction*, *12*, 439–462.

Benishek, L., & Lazzara, E. (2019). Teams in a new era: Some considerations and implications. *Frontiers in Psychology*, *10*.

Capiola, A., Alarcon, G. M., Lyons, J. B., Ryan, T. J., & Schneider, T. R. (2019). Collective efficacy as a mediator of the trustworthiness–performance

relationship in computer-mediated team-based contexts. *The Journal of Psychology*, *153*, 732–757.

Capiola, A., Baxler, H., Pfahler, M., Calhoun, C., & Bobko, P. (2020). Swift trust in ad hoc teams: A cognitive task analysis of intelligence operators in multi-domain command and control contexts. *Journal of Cognitive Engineering and Decision Making*, *14*, 218–241.

Chen, J. Y., & Barnes, M. J. (2014). Human-agent teaming for multirobot control: A review of human factors issues. *Transactions on Human-machine Systems*, *44*, 13–29.

Congressional Research Service. (2020). Joint all-domain command and control (JADC2). *In Focus [IF]*, *11493*.

Corbitt, G., Gardiner, L. R., & Wright, L. K. (2004). A comparison of team developmental stages, trust and performance for virtual versus face-to-face teams. *Proceedings of the 37th Annual Hawaii International Conference on Systems Science*.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2006). How do automation false alarms and misses affect operator compliance and reliance? *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*, 25–29.

Dixon, S. R., Wickens, C. D., & McCarley, J. S. (2007). On the independence of compliance and reliance: Are automation false alarms worse than misses? *Human Factors*, *49*, 564–572.

Fox, E. L., Houpt, J. W., & Tsang, P. S. (2021). Derivation and demonstration of a new metric for multitasking performance. *Human Factors*, *63*(5), 833–853.

Grimes, M. G., Schuetzler, R. M., & Giboney, J. S. (2021). Mental models and expectation violations on conversational ai interactions. *Decision Support Systems*, *144*.

Kanawattanachai, P., & Yoo, Y. (2002). Dynamic nature of trust in virtual teams. *Sprouts: Working Papers on Information Environments, Systems and Organizations*, *2*, 42–58.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). Soar: An architecture for general intelligence. *Artificial Intelligence*, *33*, 1–64.

Lovett, M. C., Daily, L. Z., & Reder, L. M. (2000). A source activation theory of working memory: Cross-task prediction of performance in act-r. *Cognitive Systems Research*, *1*, 99–118.

Mathieu, J., Maynard, M. T., Rapp, T., & Gilson, L. (2008). Team effectiveness 1997-2007: A review of recent advancements and a glimpse into the future. *Journal of Management*, *34*, 410–476.

McNeese, N., Demir, M., Chiou, E., Cooke, N., & Yanikian, G. (2019). Understanding the role of trust in human-autonomy teaming. *Proceedings of the 52nd Annual Hawaii International Conference on System Sciences*.

Morrison-Smith, S., & Ruiz, J. (2020). Challenges and barriers in virtual teams: A literature review. *SN Applied Science*, *2*, 1–33.

Nikolaidis, S., Hsu, D., & Srinivasa, S. (2017). Human-robot mutual adaptation in collaborative tasks: Models and experiments. *The International Journal of Robotics Research*, *36*, 618–634.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive psychology*, *7*(1), 44–64.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*, 381–410.

Priebe, M., Ligor, D. C., McClintock, B., Spirtas, M., Schwindt, K., Lee, C., Rhoades, A. L., Eaton, D., Hodgson, Q. E., & Rooney, B. (2020). *Multiple dilemmas: Challenges and options for all-domain command and control*. RAND Coorperation.

Rice, S., & McCarley, J. S. (2011). Effects of response bias and judgment framing on operator use of an automated aid in a target detection task. *Journal of Experimental Psychology: Applied*, *17*, 320–331.

Rickel, J., & Johnson, W. L. (1999). Virtual humans for team training in virtual reality. *Proceedings of the Ninth World Conference on AI in Education*.

Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. (2007). Cognitive tutor: Applied research in mathematics education. *Psychonomic Bulletin Review*, *14*, 249–255.

Saint John, A., & Levine, B. (2005). Supporting p2p gaming when players have heterogeneous resources. *Proceedings of the international workshop on Network and operating systems support for digital audio and video - NOSSDAV '05*.

Wickens, C. (1984). Processing resources and attention, varieties of attention. *R. Parasuraman and D. Davis, Eds. Academic Press*.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical issues in ergonomics science*, *3*(2), 159–177.

Zhang, H., & Houpt, J. W. (2020). Exaggerated prevalence effect with the explicit prevalence information: The description-experience gap in visual search. *Attention, Perception, & Psychophysics*, *82*(7), 3340–3356.