

Measuring Temporal Awareness for Human-Aware AI

Proceedings of the Human Factors and Ergonomics Society Annual Meeting 2023, Vol. 67(1) 1817–1823
Copyright © 2023 Human Factors and Ergonomics Society
DOI: 10.1177/21695067231192635
journals.sagepub.com/home/pro



Margaret A. Gray¹, Zhuorui Yong¹, Abhijan Wasti¹,
Esa M. Rantanen¹, and Jamison R. Heard¹

Abstract

This research investigated human performance in response to task demands that may be used to convey information about the human to an artificial agent. We performed an experiment with a dynamic time-sharing task to investigate participants' development of temporal awareness of the task event unfolding in time. Temporal awareness as an extension, or a special case, of situation awareness, may provide for useful measures of covert mental models applicable to numerous tasks and for input to human-aware AI agents. Temporal awareness measures may be used to classify human performance into the control modes in the contextual control model (COCOM): scrambled, opportunistic, tactical, and strategic. Twenty-one participants participated in a within-subjects experiment with an abstract task of resetting four independent timers within their respective windows of opportunity. The results show that temporal measures of task performance are sensitive to changes in task disruptions and difficulty and therefore have promise for human-aware AI.

Keywords

Situation Awareness, Temporal Awareness, Contextual Control, Human-Aware AI

Introduction

Human-aware artificial intelligence (AI) refers to autonomous systems that can effectively interact, collaborate, and team with humans for a variety of tasks (Kambhampati, 2020). Human-robot collaboration is a primary area for development of human-aware robots (Kumar, Arora, & Sahin, 2019; Singh & Heard, 2022), but to make any kind of AI-driven automation truly human-aware, there must be a way for the automated agents to sense various aspects of human behavior and performance to adapt accordingly and to offer a truly collaborative experience to the human. Human sensing has a long history, spanning the dead man's switches in electric streetcars and subway trains from the last century to increasingly sophisticated driver monitoring systems in highly automated automobiles in the present day (Hecht et al., 2019). However, it may be argued that human sensing research is lagging behind accelerating development of machine learning (ML) and AI-driven automation applications in all areas of life, not only in human-robot collaboration in industrial settings or in self-driving cars.

Several criteria may be developed for human sensing systems. They should be unobtrusive so as not to interfere with the human task performance in any way. They should not rely on wearable instrumentation requiring lengthy set-up or restricting human motions in task performance. They should be applicable to a wide variety of human-automation interactions (HAI) (Kaber, 2018) or human-autonomy teaming

(HAT) (O'Neill, McNeese, Barron, & Schelble, 2022). Finally, they should account for a wide variety of cognitive styles, strategies, and individual differences in humans (Feigh, 2011).

Ubiquitous and multiple-dynamic (Reason, 1990) interactions between two fundamentally different agents, humans as analog beings (Norman, 1998) and digital computers, each relying on imperfect and different but interactively and dynamically shaped models of each other, present additional research problems (Begerowski, Hedrick, Waldherr, Mears, & Shuffler, 2023; O'Neill, Flathmann, McNeese, & Salas, 2023; O'Neill et al., 2022; Stowers, Brady, MacLellan, Wohleber, & Salas, 2021). AI is trained by experience with human interactions, but these interactions are also influenced by the human experience with the AI agent. AI systems model humans as biological neural nets being trained by the systems themselves (Christian, 2020). Models and methods to enable the integration of humans and technologies in ways that optimally utilize their key strengths are needed as well (Hagenow et al.,

¹Rochester Institute of Technology, Rochester, NY, USA

Corresponding Author:

Esa M. Rantanen, Rochester Institute of Technology, 18 Lomb Memorial Dr, Rochester, NY 14623-5603, USA.
Email: esa.rantanen@rit.edu

2021a, 2021b; Pearce, Mutlu, Shah, & Radwin, 2018; Schoen, Henrichs, Strohkirch, & Mutlu, 2020).

In this paper we describe an experimental paradigm to investigate human performance and behavioral indices in dynamic tasks. Our research is based on two theoretical frameworks, situation awareness (SA) (Endsley, 1988) and the contextual control model (COCOM) (Hollnagel, 1993).

Temporal Awareness

Dynamic systems refer to environments where one must keep track of and respond to multiple changing variables. Successful control of dynamic systems implies that the users have a “mental model” of the system, allowing the user to predict system behavior and the consequences of their inputs to it. Time is an integral dimension of these dynamic systems, and is an inherent component and constraint in nearly every human activity. Having good awareness of events unfolding in time, or good temporal awareness, is crucial in creating an effective mental models of dynamic systems.

Situation awareness (SA) represents a specific, dynamic, aspect of mental models. The nearly universally accepted definition of SA by Endsley (1988) involves three levels: (1) perception of the elements in the environment, (2) comprehension of their meaning, and (3) their projection into the future. Endsley (2000) also made a distinction between mental models as representative of static knowledge about a system, whereas SA embodies a situation model, which is an extraction of time- and event-specific information from the underlying mental model. As a theoretical construct, SA has proved to be somewhat elusive, defying attempts to postulate plausible mechanisms behind it, and even its quantification in various settings. Time as a variable common to systems’ dynamics and human performance may be used as means to quantify SA.

Temporal awareness as an extension, or a special case of SA, may provide for useful measures of covert mental models applicable to numerous tasks. Appropriate task prioritization is a key performance metric in many tasks. Task prioritization further depends on accurate estimation of three temporal task parameters: (1) the time when the task becomes “available”, or the time when a window of opportunity (WO) to perform it opens, (2) the latest time by which the task must be completed, or the closing of the WO, and (3) the time required to perform the task (Rantanen, 2009).

Contextual Control Model

The COCOM model developed by Hollnagel (1993, 1998) identified several parameters that may yield useful and practical measures of operator performance. This model distinguishes four control modes: scrambled, opportunistic, tactical, and strategic. In the scrambled mode, human performance is haphazard and unpredictable, without planning, and can be best described as a state of momentary panic,

representing a complete loss of SA. The opportunistic mode is only slightly better in terms of performance or SA; the operator merely responds to the most salient events (e.g., alarms) but is not able to plan actions or predict their consequences. The tactical control mode involves planning and the operator is in control of the situation or the system, implying a moderately good SA. Finally, in strategic control mode the operator is in complete control of the task, able to consider the global context, and exhibiting good SA. Human performance in the first two modes may be characterized as reactive and in the latter two modes as proactive. Reactive and proactive behavior may be distinguishable in the timing of actions, offering a potential means for performance measurement.

The Experimental Paradigm

The experimental task in this research is an abstract time-sharing task, originally developed by Rantanen (Levinthal & Rantanen, 2004; Rantanen & Levinthal, 2005) to study workload in dynamic task settings and used in other research since (Kulomäki, Oksama, Rantanen, & Hyönä, 2022). The task is performance-dependent; speed and accuracy of performance in one trial affect the onset of subsequent trials. In this way, the task mimics the dynamic nature of real-world scenarios. The measures derived from participant responses to the task demands have been shown to be sensitive to time pressure, measured as the ratio of time required to perform a task to time available to do so (Levinthal & Rantanen, 2004) and applicable to other tasks as well (Rantanen, 2009; Rantanen & Levinthal, 2005). The current study was the first in a planned series of experiments to further develop and test human measures in a variety of tasks that could be used as input in human-aware AI and adaptive automation applications.

From a task timeline, several measures of temporal awareness may be derived, such as the proper prioritization of tasks and the “timeliness” of performance. In particular, it may be possible to measure the elapsed time from opening of a window of opportunity on individual tasks to an observable action on that task; good temporal awareness is manifested in timely performance on tasks, or consistently short “time to first action” from the opening of the window. Degradation of temporal awareness in turn is manifested in increasing variability in attending to tasks and late performance (completion of tasks after closing of the window of opportunity).

Purpose of the Research and Hypotheses

The purpose of this study was to investigate potential measures for creating representations of human temporal awareness that may provide useful feedback to human-autonomy teaming systems. The following hypotheses were developed:

H1 In an “easy” condition, participants should learn the regular pattern, or sequence, of several subtasks, having good performance indicated by little variation in performing the tasks relative to their respective WOs.

H2 As the condition changes (surprising the participants) and the (hypothetically) learned subtask performance pattern is disrupted, the participants’ performance should become poor, indicated by increased variability in performing the subtasks relative to their respective WOs.

H3 The participants will learn the new pattern in time, again exhibiting little variation in performing the subtasks relative to their respective WOs.

Method

Participants

Twenty-one students recruited from an undergraduate psychology course volunteered to participate in exchange for course credit. This research was approved by the Office of Human Subject Research at the Rochester Institute of Technology, and all participants gave their informed consent to participate. The participants were 18 to 21 years of age ($M = 18.8$ years); 11 self-identified as female, 8 as male, and 2 as non-binary. The participants had different ethnicities including Hispanic, White, Black, and Asian. All participants had the highest completed education of high school or equivalent. Twenty participants were native English speakers; one participant was deaf, and one hard of hearing. All but six participants were avid players of video games.

Apparatus

A software program, developed on PsychoPy and Python3, simulated a time-sharing task. A computer screen was divided into four panes, which were masked. To unmask a pane and reveal a subtask in it, the participant had to move a cursor to that pane. Once the cursor was moved to another pane, the previous pane was again masked and a different subtask revealed in the pane where the cursor was. The subtasks consisted of progress bars, which were to be reset by typing a 4-digit code within their respective WOs, indicated on the bars. Moreover, when the WO opened the bar turned green and when the WO closed without reset code entered the bar turned red. The participants’ task was to monitor the status of four independent progress bars and reset them within their respective WOs (i.e., before the WO closed; a bar could, and should, be reset late, or after closing the WO). Figure 1 shows the experimental task.

Independent Variables

We designed two different experimental conditions. The first three minutes in the task were considered “easy”, with a

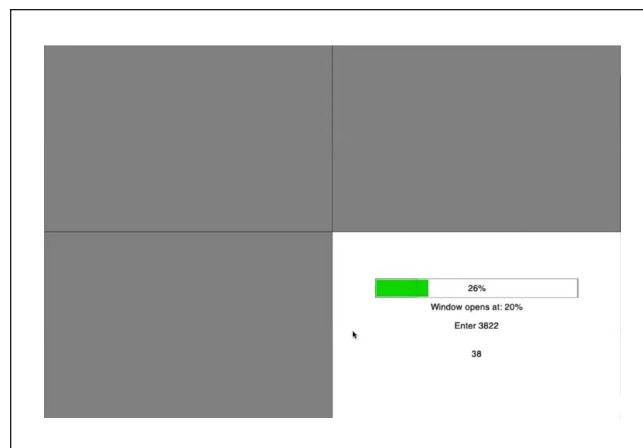


Figure 1. Participant’s view of a pane where the WO has opened at 20%. The progress bar is currently at 26%, and the participant has typed in “38” of the 4-digit reset code. Three other progress bars are independently moving in the masked panes.

relatively leisurely pace at which the bars were to be reset to allow participants develop a good temporal awareness of the subtasks (bar resets) to be performed. After three minutes the bar speed was increased and the sequence in which the bars were to be reset was changed for the second half, or another 3 minutes, of the experiment, surprising the participants. Re-learning a new sequence of subtasks and the overall faster pace would make re-acquiring temporal awareness difficult, hypothetically reflected in observable performance. Specific task parameters are presented in Table 1.

Dependent Variables

The experimental program recorded several time-stamped events and the participants’ actions. We also tracked the mouse position and keyboard events by the program. Table 2 shows the raw measures, and Table 3 shows the key events.

From these data we derived the time to first action (TFA) within each pane, calculated from the length of time between the opening of the WO to when the user typed in the first digit in the reset code. This was the primary performance metric representing the timeliness of performance in the four subtasks.

Experimental Design

This was a within-subject design. We divided the duration of the experiment into six 1-minute time epochs, the first three representing the “Easy” condition and the last three representing the “Hard” condition. We wanted to examine learning of the task during the first half of the experiment and the effect of quickening the overall task pace and disruption of learned patterns on task performance.

Table 1. Subtask Parameters to Create Two Experimental Conditions, “Easy” and “Hard”.

Settings	Cond.	Pane 1	Pane 2	Pane 3	Pane 4
Bar Duration	Easy	25s	45s	30s	55s
	Hard	20s	30s	15s	20s
Window Opening	Easy	50%	50%	60%	20%
	Hard	40%	45%	40%	60%
Window Duration	Easy	12.5s	22.5s	12s	44s
	Hard	12s	16.5s	9s	8s
Bar Speed	Easy	4.00%/s	2.22%/s	3.33%/s	1.82%/s
	Hard	5.00%/s	3.33%/s	6.67%/s	5.00%/s

Table 2. Raw Data Output and descriptions.

Measurable	Description
Mouse position	The current X and Y coordinates of the cursor on the screen, between -1.000 and 1.000
Bar progress	For each pane, the current position of the progress bar, recorded as a number between 0.00 and 100.00
Pane keystrokes	For each pane, the current keystrokes that have been entered
Mouse-in-pane	Which pane the cursor is currently in
Event-in-pane	Which pane the event occurred in

Table 3. Key Events that triggered recording a timestamp.

Event	Definition
1.Switch Panel	Cursor switched into a different pane
2.Key Stroke	User typed in a single keystroke
3.WO-Open	The WO opened within an individual pane
4.WO-Close	The WO closed within an individual pane
5. Pane reset	User entered the correct input, the progress bar goes back to 0, and a new question is assigned

Procedure

The researchers briefly introduced the study to the participants. First, participants completed a demographic information survey, followed by the researchers explaining the task and how to interact with the program. Participants read, agreed, and signed the consent form and started the experiment when ready. The experiment lasted a total of 6 minutes. After the experiment program was finished, researchers conducted a brief one-on-one interview with participants to identify what changes they noticed, their strategies to handle the task, and how they evaluated their own performance.

Results

Each minute within the six-minute experimental trial was split up into 1-minute “epochs”, with the “Easy” condition

being divided into epochs 1–3, and the “Hard” condition into epochs 4–6. For each epoch, the TFA values were not normally distributed, as is common with timing data, and as shown in Figure 2. Log transformations worked well to normalize the data. Figure 3 shows the mean and 95% confidence interval for TFA for each of the six epochs (backtransformed from the log scale), and Table 4 lists the values of all descriptive statistics on both the log and back-transformed scales.

Within the “Easy” condition of the experiment, the mean log TFA decreased from 3.22 to 2.44 to 2.34, and the standard deviation of log TFA also decreased from 1.08 to 0.99 to 0.88. Within the “Hard” condition, the mean log TFA increased from 1.24 to 1.31 from epochs 4 to 5, and then in epoch 6 decreased to 1.26. The standard deviation of log TFA decreased from 0.80 to 0.78 to 0.73.

Between the “Easy” and “Hard” conditions, from epochs 3 to 4, the mean log TFA increased by 0.49, from 0.85 to 1.24. The standard deviation of the log TFA decreased from 0.88 to 0.80. A two-sample Wilcoxon rank test (Mann-Whitney U test) was performed to compare the two conditions (epochs 1-3 vs epochs 4-6), and produced a p-value of less than 0.0001.

These trends support our hypotheses. Within the first three minutes (epochs) and in the “Easy” condition, participants clearly learned the task and improved in their performance, resetting the bars soon after opening the WO (decreasing mean log TFA) and exhibiting increasing consistency in their performance (decreasing standard deviation of log TFA). Interruption of this good performance by increasing the task pace and disrupting the learned reset pattern resulted in poorer performance in epoch 4 (lower mean log TFA), but decreased variability (decrease in standard deviation of log TFA). Although the performance appears to have been stabilized in epochs 5 and 6, it never reached the same low level achieved in epoch 3 within the remaining experimental duration.

Discussion

Within the “Easy” condition, the decrease in mean log TFA and the decrease in standard deviation of log TFA provides

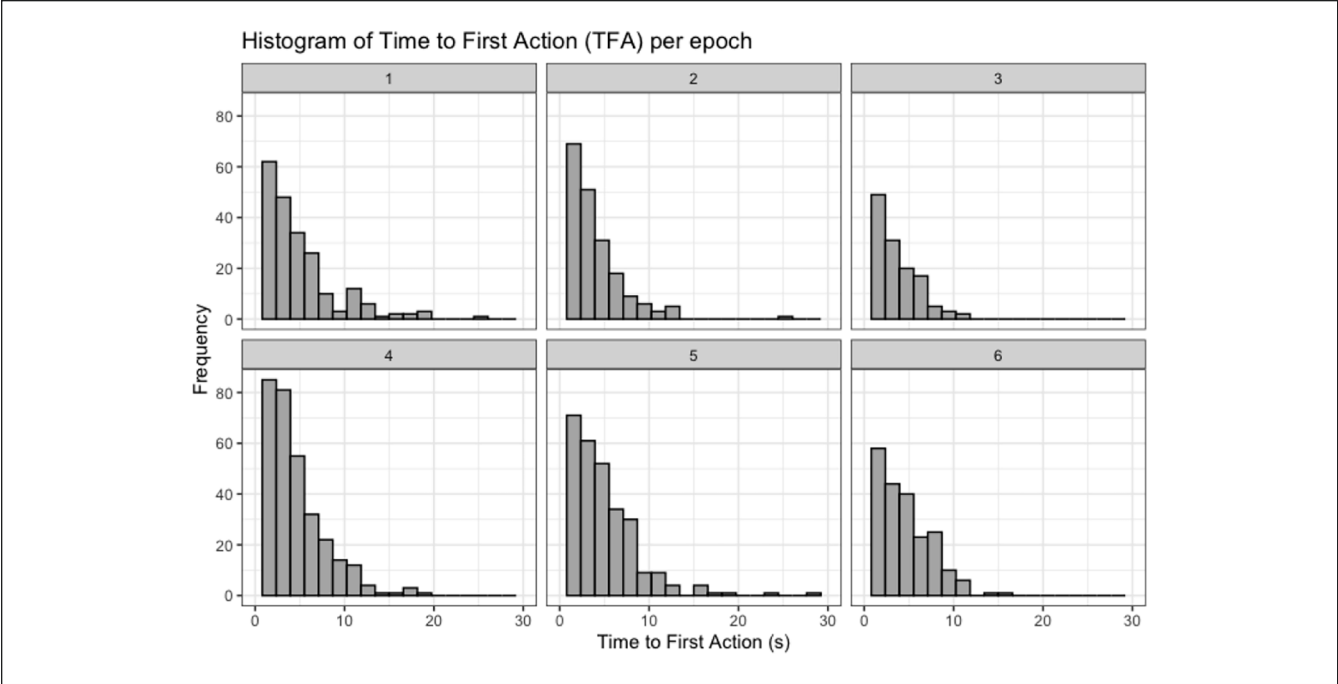


Figure 2. Histograms showing distributions of participants' time to first action (in seconds), separated by epoch. The first three epochs represent each minute under the easy condition, and the last three represent each minute under the difficult condition. Log transformations worked well to normalize data for all epochs.

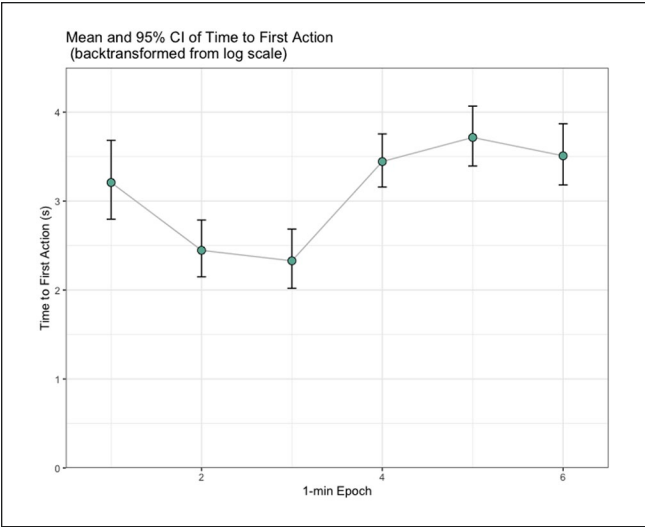


Figure 3. Mean and 95% CI (backtransformed from log scale) for time to first action (TFA) (in seconds), for each of the six epochs.

evidence to support our H1 hypothesis that participants will improve their performance condition as they spend time getting used to the pattern. Within the “Hard” condition, it is interesting that there does not seem to be much of a significant decline in mean log TFA or a decrease in the standard

Table 4. Summary Statistics For TFA For Each Epoch. Log mean and log standard deviation are on the log scale, mean and 95% CI have been backtransformed from that scale.

Epoch	N	log mean	log stdev	mean	95% CI
1	237	1.17	1.08	3.22	[2.81, 3.70]
2	222	0.89	0.99	2.44	[2.12, 2.79]
3	148	0.85	0.88	2.34	[2.04, 2.68]
4	326	1.24	0.80	3.46	[3.20, 3.74]
5	287	1.31	0.78	3.71	[3.36, 4.09]
6	215	1.26	0.73	3.53	[3.20, 3.89]

deviation of mean log TFA. We had originally hypothesized (H3) that, as participants learned the new pattern, they would display a similar trend of a decreasing mean log TFA and decreasing standard deviation of log TFA. Our findings do not support that hypothesis.

Between the “Easy” and “Hard” conditions, there was a jump in TFA, suggesting that the change in conditions did impact performance. A Mann-Whitney U-test showed a significant difference between the two conditions ($p < 0.0001$). We hypothesized (H2) that, as the condition changes, the reset pattern is disrupted, and variability in the participants’ performance would increase. The first part of our hypothesis is supported, but the variability decreased, so the second part of the hypothesis was not supported.

There could be several explanations for why there is not much of a declining trend in the “Hard” condition, or why the TFA for epoch 4 is slightly lower than the TFA in epochs 5 and 6: Perhaps three minutes is not enough time for the participants to fully learn the new pattern, or perhaps in epoch 4 participants were able to take a “break” while all of the panes simultaneously reset, and thus had more downtime before responding to the WO openings.

In any case, our results show that temporal measures of task performance are sensitive to changes in task disruptions and difficulty. There are certainly many more temporal measures to be derived from the raw data output our experimental paradigm can provide, offering a more complete view to the development and changes in human temporal awareness in dynamic tasks, and potentially providing input to human-aware AI agents for their adaptation to human performance.

Limitations

Due to the homogeneity of our participant population, this experiment should be repeated with a larger, more heterogeneous sample. The program output of raw data was also not conducive for derivation of human performance variables. Therefore, development of standard of data output from human-automation interactions for reliable and real-time calculation of valid human performance measures is a critical task for further experiments.

Future Research

This pilot study showed the feasibility of our experimental paradigm to study human performance in a dynamic task through data unintrusively derived from the task performance. Our experimental program will allow for many additional measures to be recorded, including mouse positions, pane switching before the WO opens (system monitoring), and other timed actions relative to the task. Additionally, there is a good amount of flexibility we have in defining our experimental conditions. Having a longer period per experimental conditions may help us identify trends more clearly and would allow us to have a better view of when someone settles into a pattern. It also could be interesting to incorporate “breaks” into the experiment to measure participants’ situation awareness in different ways, for example, by SAGAT (Endsley, 1988), or to incorporate SPAM (Durso & Dattel, 2004) into the task.

The next steps in this research program will develop an array of measures to be derived from both humans (behavioral, psychophysiological, and subjective) and the system (temporal and cognitive demands of the tasks), and use of machine learning for data fusion (system and human measures) and in classification of human performance into the COCOM modes (scrambled, opportunistic, tactical, and strategic). We will also research the transparency of the ML algorithms and robustness of the COCOM classifications

across diverse user populations and task demands for their suitability for human-aware AI.

Acknowledgments

This research was supported by the National Science Foundation under Award No. DGE-2125362. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

ORCID iDs

Zhuorui Yong  <https://orcid.org/0009-0001-4237-8357>

Esa M. Rantanen  <https://orcid.org/0000-0001-9666-4458>

References

- Begerowski, S. R., Hedrick, K. N., Waldherr, F., Mears, L., & Shuffler, M. L. (2023). The forgotten teammate: Considering the labor perspective in human-autonomy teams. *Computers in Human Behavior*, 145, 107763.
- Christian, B. (2020). *The alignment problem: Machine learning and human values*. WW Norton & Company.
- Durso, F. T., & Dattel, A. R. (2004). SPAM: The real-time assessment of SA. In S. Banbury & S. Tremblay (Eds.), *A cognitive approach to situation awareness: Theory and application* (pp. 137–154). Ashgate.
- Endsley, M. R. (1988). *SAGAT: A methodology for the measurement of situation awareness* (Tech. Rep. No. NOR DOC 87-83). Hawthorne, CA: Northrop Corp.
- Endsley, M. R. (2000). Situation models: An avenue to the modeling of mental models. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 44, pp. 61–64).
- Feigh, K. M. (2011). Incorporating multiple patterns of activity into the design of cognitive work support systems. *Cognition, Technology & Work*, 13(4), 259–279. Retrieved from <https://doi.org/10.1007/s10111-010-0165-0> doi: 10.1007/s10111-010-0165-0
- Hagenow, M., Senft, E., Radwin, R., Gleicher, M., Mutlu, B., & Zinn, M. (2021a). Corrective shared autonomy for addressing task variability. *IEEE robotics and automation letters*, 6(2), 3720–3727.
- Hagenow, M., Senft, E., Radwin, R., Gleicher, M., Mutlu, B., & Zinn, M. (2021b). Informing real-time corrections in corrective shared autonomy through expert demonstrations. *IEEE Robotics and Automation Letters*, 6(4), 6442–6449.
- Hecht, T., Feldhütter, A., Radlmayr, J., Nakano, Y., Miki, Y., Henle, C., & Bengler, K. (2019). A review of driver state monitoring systems in the context of automated driving. In *Proceedings of the 20th congress of the international ergonomics association (iea 2018) volume vi: Transport ergonomics and human factors (tehf), aerospace human factors and ergonomics 20* (pp. 398–408).
- Hollnagel, E. (1993). *Human reliability analysis: Context and control*. Academic press.
- Hollnagel, E. (1998). *Cognitive Reliability and Error Analysis Method (CREAM)*. Elsevier.
- Kaber, D. B. (2018). Issues in human–automation interaction modeling: Presumptive aspects of frameworks of types and levels

- of automation. *Journal of Cognitive Engineering and Decision Making*, 12(1), 7–24.
- Kambhampati, S. (2020). Challenges of human-aware ai systems: Aai presidential address. *AI Magazine*, 41(3), 3–17.
- Kulomäki, J., Oksama, L., Rantanen, E., & Hyönä, J. (2022). Attention control in a demanding dynamic time-sharing environment: An eye-tracking study. *Attention, Perception & Psychophysics*, 84(2), 352–371.
- Kumar, S., Arora, S., & Sahin, F. (2019). Speed and separation monitoring using on-robot time-of-flight laser-ranging sensor arrays. In *IEEE 15th International Conference on Automation Science and Engineering (CASE)* (pp. 1684–1691).
- Levinthal, B. R., & Rantanen, E. M. (2004). Measurement of task-load and performance in a dynamic multi-task experiment. In *Proceedings of the 48th Human Factors and Ergonomics Society Annual Meeting* (Vol. 48, pp. 567–570).
- Norman, D. (1998). *The invisible computer*. The MIT Press.
- O'Neill, T. A., Flathmann, C., McNeese, N. J., & Salas, E. (2023). Humanautonomy teaming: Need for a guiding team-based framework? *Computers in Human Behavior*, 146, 107762.
- O'Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human– autonomy teaming: A review and analysis of the empirical literature. *Human factors*, 64(5), 904–938.
- Pearce, M., Mutlu, B., Shah, J., & Radwin, R. (2018). Optimizing makespan and ergonomics in integrating collaborative robots into manufacturing processes. *IEEE transactions on automation science and engineering*, 15(4), 1772–1784.
- Rantanen, E. M. (2009). Measurement of temporal awareness in air traffic control. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 53, pp. 6–10).
- Rantanen, E. M., & Levinthal, B. R. (2005). Time-based modeling of human performance. In *Proceedings of the 49th Human Factors and Ergonomics Society Annual Meeting* (Vol. 49, pp. 1200–1204).
- Reason, J. (1990). *Human error*. Cambridge University Press.
- Schoen, A., Henrichs, C., Strohkirch, M., & Mutlu, B. (2020). Authr: A task authoring environment for human-robot teams. In *Proceedings of the 33rd annual acm symposium on user interface software and technology* (pp. 1194–1208).
- Singh, S., & Heard, J. (2022). Human-aware reinforcement learning for adaptive human robot teaming. In *17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)* (p. 1049-1052). doi: 10.1109/HRI53351.2022.9889530
- Stowers, K., Brady, L. L., MacLellan, C., Wohleber, R., & Salas, E. (2021). Improving teamwork competencies in human-machine teams: Perspectives from team science. *Frontiers in Psychology*, 12, 590290.