The Role of Autonomy Levels and Contextual Risk in Designing Safer AI Teammates

Allyson I. Hauptman
School of Computing
Clemson University
Clemson, SC 29631
Email: ahauptm@clemson.edu

Beau G. Schelble
School of Computing
Clemson University
Clemson, SC 29631
Email: bschelb@clemson.edu

Christopher Flathmann
School of Computing
Clemson University
Clemson, SC 29631
Email: cflathm@clemson.edu

Nathan J. McNeese
School of Computing
Clemson University
Clemson, SC 29631
Email: mcneese@clemson.edu

2024 IEEE 4th International Conference on Human-Machine Systems (ICHMS) | 979-8-3503-1579-0/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICHMS59971.2024.10555844

***Abstract*—As AI becomes more intelligent and autonomous, the concept of human-AI teaming has become more realistic and attractive. Despite the promises of AI teammates, human-AI teams face new, unique challenges. One such challenge is the declining ability of human team members to detect and respond to AI failures as they become further removed from the AI's decision-making loop. In this study, we conducted virtual experiments with twelve experts in two different teaming contexts, cyber incident response and medical triage, to understand how contextual risk impacts human teammate situational awareness and failure performance over a human-AI team's action cycle. Our results indicate that situational awareness is more closely tied to context, while failure performance is more closely tied to the team's action cycle. These results provide the foundation for future research into using contextual risk in determining optimal autonomy levels for AI teammates.**

I. Introduction

Advancements in artificial intelligence (AI) technologies have led to the possibility of AI agents taking on full-scale team roles as part of a human-AI team [1], [2]. These human-AI teams can lead not only to greater efficiency and productivity but also allow them to complete tasks that would be impossible without AI [3]–[6]. These new hybrid teams are beginning to be fielded in a variety of areas, such as education [7], eSports [5], [8], and law enforcement [9]. Despite their advantages, these hybrid teams also present unique challenges for humans on the team, such as how to convey an AI agent's decision-making logic and capabilities to human team members [10] and facilitate the generation of trust between human and artificial teammates [5], [11].

The amount of input humans have in an AI agent's decision-making loop can be referred to as its level of autonomy (LOA) [12]. As the agent's LOA increases, humans become further removed from the AI's decisions, which consequently causes a decrease in their situational awareness of the AI's actions and operations, a phenomenon known as the ironies of autonomy [13]. In fact, throughout an AI's action cycle - Information Selection, Information Analysis, Action Selection, Action Implementation - its autonomy functionally increases and causes this issue to worsen [14]. The extent to which these issues of declining situational awareness exist in applied contexts has been debated amongst researchers, particularly in regards to how generalizable it is to human-AI collaboration contexts and whether AI designers should err on the side of more or less AI autonomy [15]–[17]. In response, some recent human-AI teaming research has suggested that the reason for this discrepancy is that different human-AI collaboration contexts incur various levels of risk; thus, the degree to which the ironies of autonomy should be considered in an AI teammate's design should be affected by the context in which the AI will operate [18].

In considering the debate over the possibilities versus the consequences of high AI autonomy, researchers and designers have to take into account numerous aspects of the AI agent's purpose. For human-AI teams, this includes not only its technical capabilities but also how it operates as a *teammate* for its human collaborators and the effects it has on their abilities to perform their own duties [1], [5]. The need to understand how to design AI teammates that can enhance teamwork while mitigating negative consequences of higher AI autonomy motivates this study and the following research questions:

RQ1: How does a human teammate's situational awareness, failure performance, and risk perceptions change over an AI agent's action cycle during a team task?

RQ2: How do these changes differ between contexts of different risk types?

We conducted virtual experiments with twelve professionals across two contexts with varying risk types, in which the participants completed a routine team task with an AI teammate and reported their changes in situational awareness, failure performance, and perceived risk levels. The results of this study show the importance of context in designing AI teammates and serve as a starting point for AI designers to conduct future research on how to define and map types of contextual risk to AI autonomy levels for human-AI teams.

II. Background

An AI agent can be programmed to operate with one or more autonomy levels. Based on the Levels of Automation [19], O'Neill and colleagues developed an adapted LOA scale for AI autonomy, which posits that an agent can be considered autonomous at and beyond Level 5 [12]. At and beyond level

979-8-3503-1579-0/24/$31.00 ©2024 IEEE

five, the intelligent system takes on more autonomy, such that it is taking on more roles and responsibilities and even the authority to execute decisions [20], [21]. This then reduces the cognitive effort the human requires within the decision-making process. As these systems become more intelligent, they need less human oversight [22]. In levels six to ten, the intelligent system is not only reducing the degree of notification it uses to inform the user, but it is also decreasing the amount of awareness the human has in its decision-making and its progress in completing its assigned tasks [22], [23]. At the final level, level ten, the intelligent system does not need to factor in the human to any degree and independently acts with complete autonomy.

In addition to the position of the human in an AI agent's decision-making loop, the actions that the AI is programmed to complete also play a role in classifying its autonomy. Wickens and colleagues referred to this as the agent's Degree of Autonomy (DOA), which describes how the agent's autonomy functionally increases as it operates throughout the action cycle [14]. What is most important about the existence of DOAs is that as an agent's DOA increases, research shows that the situational awareness of the humans with whom the AI interacts decreases [14]. Situational awareness can be viewed as a three-tier concept where an entity perceives and considers individual factors, task factors, and environmental factors to make decisions and perform actions [24]. As an agent's autonomy increases, humans depend more on the AI to complete tasks they cannot complete themselves [25]. As a result, when an AI agent fails or encounters a situation it does not know how to consider, humans may be unable to prevent or reverse the impacts of that failure [25].

This issue is exacerbated in human-AI teams, where AI agents take on complete, independent team roles [12]. Teammates require a minimum level of situational awareness of their teammates' decisions and actions to perceive and adapt to changes [23], [26]. Through the processes of teamwork, members of the team are expected to aid in the group's shared awareness of their progress and changes to the team's environment and goals, a shared awareness that can help in responding to any failures or unforeseen events [27], [28]. However, AI teammates are most lucrative to teams because of the AI's potential to perform tasks differently and beyond human capabilities [4], [5]. As such, even if human teammates detected an AI failure, they would require a certain amount of time to understand the scope of the failure and how to respond to it.

This issue of a human's declining ability to detect and respond to AI failures as its autonomy increases is known as the Lumberjack Effect, which refers to the higher the tree (AI autonomy), the harder the fall when it encounters an axe (a failure state) [14]. This concept has been somewhat debated due to its implications for AI design. Researchers have cautioned that the extent to which the Lumberjack Effect exists in applied contexts varies and that the benefits of higher AI autonomy should not be abandoned due to fears over the possible consequences [15], [16]. Human-AI teaming research has since emphasized that the context in which a human-AI team operates and the risks associated with an AI agent's tasks are essential factors to consider in selecting an AI agent's optimal autonomy levels [1], [18]. Thus, this study is motivated by understanding how these factors, context and risk level, affect human teammate situational awareness and the subsequent effect on team failure performance.

III. Methods

To assess how context and AI autonomy levels affect human teammate perceptions, we designed two virtual experiments, one for each context. This method was appropriate, because using human-AI teams in applied contexts is still in its infancy, so real-world experiments would not have been possible. Instead, we created realistic scenarios virtually that allowed our participants to perform the same tasks they do daily but with the addition of an AI teammate. The simulations also allowed the researchers to purposefully introduce the AI failures necessary to obtain repeated measures of the participants' changing capabilities and perceptions throughout the task.

We selected cyber incident response and emergency room triage for the specific risk contexts. There are multiple reasons why these two contexts were appropriate for the method and the research questions. First, both contexts are currently actively experimenting with including AI teammates in their processes [29], [30]. Second, these two contexts each represent very different types of risk: virtualized data risk and kinetic human risk. The failures of cyber incident response teams for small business networks primarily affect data availability and integrity [31]. In contrast, the shortcomings of medical triage teams directly impact the health and welfare of human patients [32]. These risk types are different enough to assess if there are significant differences in how contextual risk affects human perceptions and performance.

A. Participants

Participants for this study were recruited through targeted emails to known professionals and snowball sampling. This targeted recruitment was done to ensure a high level of subject matter expertise, as contextual expertise is vital to collecting accurate data from contextual inquiries [33]. Six participants were recruited for each risk context. Participants in the medical triage context possessed at least eight years of experience in the medical field ($M = 18.4$). In the incident response context, participants possessed at least two years of IT or networking experience ($M = 12.5$). Additional participant demographics are shown in Table I.

B. Experiment Exercises

To replicate an AI that begins its work in a highly autonomous state, participants were told that during the exercise, the AI would be operating with an autonomy level of "informs the human only if asked," which equates to Level 8 on the LOA scale [12]. The intent of starting an AI in a highly autonomous state was to position the AI in a highly independent position

TABLE I
STUDYY PARTICIPANTS

Context	Gender	Age	Occupation
Incident Response	Male	31	Graduate Student
Incident Response	Male	31	Network Security Analyst
Incident Response	Male	30	Network Systems Engineer
Incident Response	Male	31	Loan Servicing
Incident Response	Male	34	Software Engineer
Incident Response	Male	28	Site Reliability Engineer
Medical Triage	Female	37	Medical Doctor
Medical Triage	Female	64	Registered Nurse
Medical Triage	Female	30	Registered Nurse
Medical Triage	Female	44	Graduate Nursing Student
Medical Triage	Male	64	Registered Nurse (Ret.)
Medical Triage	Male	35	Graduate Nursing Student

that would inherently cause the human's situational awareness to decline more quickly throughout a short exercise. This would be the autonomy level the participant had to keep in mind when considering their ability to reverse an AI failure, as well as the risk level during the exercise itself. This level also corresponds with what is considered in the HCI community to be a highly autonomous system [12].

1) Context 1: Cyber Incident Response: For the first context, six subject matter experts in cyber security were given a cyber incident response scenario. In this exercise, the AI teammate is the network analyst tasked with conducting a rapid network analysis of a small business network that has been the victim of phishing attacks. The human participant serves as the security officer tasked with inspecting the content of the emails that the AI flags as suspicious. The participant's task was simulated through an online phishing quiz [34], which was selected after piloting with experts over the use of PDFs of the supposed emails as more realistic and real-time. Throughout the scenario, the AI failed four times, once per phase of the action cycle:

1) Information Acquisition: The AI collects network traffic from the wrong interface.
2) Information Analysis: The AI misses key non-standard ports in its HTTP analysis.
3) Action Selection: The AI selects a PDF based on an odd but benign domain name.
4) Action Execution: The AI downloads the malicious PDF to the wrong server, which is connected to the business network.

2) Context 2: Medical Triage: For the second context, six subject matter experts in medical treatment were tasked with completing an emergency room triage scenario, which was selected based on its well-known and used triage process [35]. The AI teammate was responsible for conducting phone triage with a single patient and was represented by a prototype AI triage agent, Dasha AI [36]. The human was responsible for in-person triage, which was simulated through a quiz delivered through Qualtrics based on the National Council Licensure Examination (NCLEX) for nursing [37]. Piloting revealed that some more experienced personnel further removed from their schooling might need a refresher on these principles. Thus, a primer was included in the introduction email to the participants. During the scenario, the AI would make four consecutive failures while communicating with a patient over the phone:

1) Information Acquisition: AI misinterprets patient test results as "negative"
2) Information Analysis: AI incorrectly matches the patient's answers to a percentile of a problem
3) Action Selection: AI selects lower priority for the patient based upon failure in the previous stage
4) Action Implementation: AI teammate disconnects with the patient when should be kept on the line

C. Procedure

All participants received an introductory email that included a video explanation of the tasks that the AI teammate would complete and a link to a pre-survey with an informed consent form and demographic information. Before the participant's time slot, they received an email with the Zoom link and pre-survey for demographic information. Once participants logged into the video call, they were asked to provide consent to record, after which the researcher shared their screen with the AI agent and began the scenario. During the experiment, the researcher introduced the failures noted above for each of the four phases, after which measures for the participants' situational awareness, failure performance, and perceived risk level were collected.

D. Perceptual Measures

The perceptual measures taken throughout the experiment included situational awareness, failure performance, and perceived risk level. Each of these constructs was measured using a single item, with the item measuring situational awareness asking "How would you rate your situational awareness of your AI teammate's actions?". The item measuring failure performance asked, "How would you rate your ability to respond to and fix the AI's failure?". Lastly, the item measuring perceived risk level asked "How would you rate the risk level of your AI teammate's failure at this point in time?" Participants responded to all three items using a Likert scale ranging from 1 to 10. These constructs were measured using a single item to ensure that experimental fatigue did not set in, which is a frequently used strategy used to measure perceptual and emergent states in human-AI teaming [38].

IV. RESULTS

These contextual inquiries produced interesting data for analysis. During each of the four action stages, every participant rated three perceptions (situational awareness, failure performance, and risk level) on a 10-point scale. Although the sample size of participants is small, they were all subject matter experts and a total of 12 measures (4 per dependent variable) per participant were collected for a total of 144 measurements. These measurements were analyzed using a repeated measures (RM) ANOVA and considered the main effects of action stage, risk context and their interaction, the intent of which was to both quantify how the AI's increasing autonomy affected human perceptions and how these impacts differ between different types of applied human-AI team contexts.

A. Participant Perceptions

RM ANOVAs were conducted to examine the effect of the four action stages and context on participants' perceived situational awareness, failure performance, and risk level. Greenhouse-Geisser corrections were used to correct for a violation of sphericity, and Holm post-Hoc tests were used to explore the main effects.

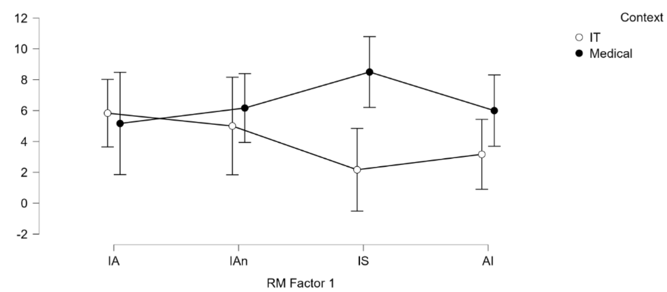

Fig. 1. Perceived situational awareness with 95 % confidence intervals.

a) Perceived Situational Awareness: The results for perceived situational awareness are shown in Figure 1. The RM ANOVA revealed a non-significant main effect for action stage (F(2.73, 27.30) $=$ 0.41, p $=$ .729, η^2 $=$ .02); however, a significant main effect for context type (F(1, 10) $=$ 5.05, p $=$ .048, $\eta^2 =$ .15) and a significant interaction effect between scenario and action stage (F(2.73, 27.30) $=$ 4.36, p $=$ .014, $\eta^2 =$ .17) were found. For the main effect of context, situation awareness was perceived to be higher in the medical context (M $=$ 6.55, SE $=$ 0.76) compared to the networking context (M $=$ 4.04, SE $=$ 0.76).

For the interaction between the context and action stage, an analysis of the simple main effect when controlling for the action stage revealed that there was a non-significant difference between contexts at the IA (F(1, 30) $=$ 0.10, p $=$ .757), IAn (F(1, 30) $=$ 0.45, p $=$ .518), and AI (F(1, 30) $=$ 3.13, p $=$ .107) action stages, but there was a significant difference in context at the IS stage (F(1, 30) $=$ 53.88, p $<$ .001). A Holm post-hoc test revealed that during the IS stage, participants in different contexts perceived significantly different situation awareness (t $=$ 3.87, SE $=$ 1.64, p $=$ .013); in particular, participants in the medical context (M $=$ 6.46, SE $=$ 0.76) perceived significantly higher situation awareness than those in the networking context (M $=$ 4.04, SE $=$ 0.76). Additionally, when controlling for context, the differences between action stages were found to be non-significant for both medical (F(3, 30) $=$ 2.03, p $=$ .152) and networking contexts (F(3, 30) $=$ 2.73, p $=$ .080).

b) Perceived Failure Performance: The results of perceived failure performance are shown in Figure 2. Only the main effect of the action stage on failure performance (F(1.77,

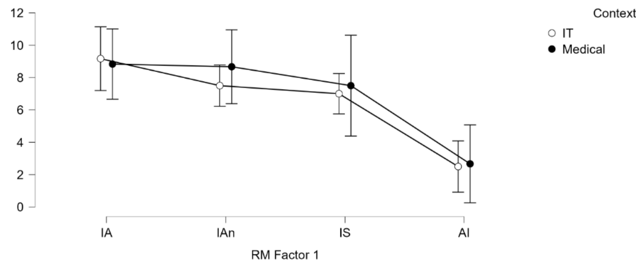

Fig. 2. Perceived failure performance with 95 % confidence intervals.

17.65) $=$ 24.60, p $<$ .001, η^2 $=$ .64) was significant, but the main effect of the context (F(1.77, 17.65) $=$ 24.60, p $=$ .719, η^2 $=$ .01) and the interaction effect (F(1, 10) $=$ 0.45, p $=$ .518, η^2 $<$ .01) were both non-significant. A Holm post-hoc test revealed that compared to the AI stage, the IS, IAn, and IA stages were all significantly different, but all other post-hoc comparisons were non-significant. In particular, perceived failure at the AI stage was lower (M $=$ 2.58, SE $=$ 0.57) than the IS (M $=$ 7.25, SE $=$ 0.57), IAn (M $=$ 8.08, SE $=$ 0.57), and IA (M $=$ 9.00, SE $=$ 0.57) stages.

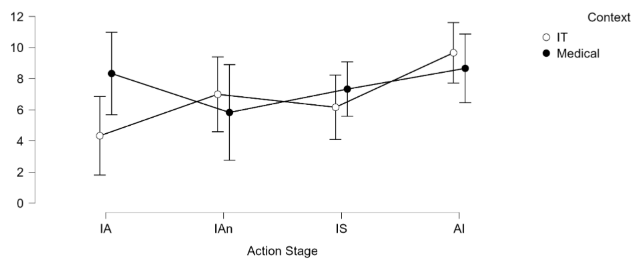

Fig. 3. Risk level of an AI failure with 95 % confidence intervals.

c) Risk Level: The results of the RM ANOVA for Risk Level revealed that the main effect of the action stage (F(2.46, 24.63) $=$ 4.29, p $=$ .02, η^2 $=$ .15) and the interaction effect between the action stage and context (F(2.46, 24.63) $=$ 3.45, p $=$ 0.04, η^2 $=$ .12) were significant, but the main effect of the context was insignificant (F(1, 10) $=$ 0.41, p $=$ .536, η^2 $=$ .02). For the main effect of the action stage, a Holm post-hoc test revealed that compared to the AI stage, the IA and IAn stage was significantly different. Specifically, participants perceived greater risk levels in the AI stage (M $=$ 9.17, SE $=$ 0.81) compared to the IA (M $=$ 6.33, SE $=$ 0.81), and IAn (M $=$ 6.42, SE $=$ 0.81).

For the interaction effect, the simple main effects revealed that when controlling for the context, there was only a significant difference between action stages in the networking context (F(3, 30) $=$ 6.43, p $=$ .005) but not the medical context (F(3, 30) $=$ 1.75, p $=$ .199). Specific analysis of the Holm post-hoc effects revealed that in the networking condition, there was a significant difference in the risk perceived by participants in the AI stage compared to the IA stage, with risk at the AI stage (M $=$ 9.67, SE $=$ 1.09) perceived to be greater than that at the IA stage (M $=$ 4.33, SE $=$ 1.09).

d) Summary of Quantitative Analysis: In sum, when we look at the trends of responses on human teammate perceptions, we can see that context played a much more significant role in impacting changes in situational awareness and risk level, while the action stage more heavily influenced failure Performance. Failure performance for both contexts dropped dramatically in the Action Implementation stage, the final stage when an AI teammate's functional autonomy would be highest. Interestingly, the trends for situational awareness and failure performance differed, which is a departure from prior human-AI collaboration research. The unique role of teaming that places increased responsibilities on an AI teammate seems to affect how closely these two variables mirror one another, and this discrepancy highlights the need for a third factor, contextual risk, in helping to triangulate the optimal autonomy levels for AI teammates.

V. DISCUSSION

The current study examined how context affected participants' perceptions of varying levels of autonomy through virtual experiments. The findings from the study's analysis address the two research questions and provide helpful information to practitioners of human-AI teaming. Specifically, RQ1 sought to understand how a human teammate's perception of situational awareness, risk awareness, and failure performance change due over the AI teammate's action cycle. The results found that while a teammate's failure performance and perceived risk levels were closely tied to AI's increasing autonomy over the action cycle, situational awareness was not. RQ2, on the other hand, examined whether or not the effect of the AI teammate's increasing autonomy level on the human teammate changed based on the context where they work. The study showed that the context the team was operating within, IT or medical, significantly changed the participants' responses to their situational awareness measure. In the IT context, participants had increasing situational awareness up until the final stage of the action cycle. In contrast, in the medical context, participants' situational awareness declined until the final stage of the action cycle.

The results of this study provide exciting insight into our research questions. The trends in our data show that specific human factors such as situational awareness, failure performance, and perceived risk level are affected unequally by the action cycle and an AI teammate's functional autonomy level. While situational awareness dramatically declined from the information analysis to the action selection stage in the incident response context, it increased in the medical context. This increase is a departure from the expected decline that previous research claims should occur [14]. It is interesting that in this study, some teammate perceptions were more linked to context than the action stage. While this aligns with empirical testing of failure performance and the ironies of autonomy [14], it also shows that an AI agent's role in the team has a more significant effect on a teammate's situational awareness.

In contrast to situational awareness, a human teammate's ability to correct an AI failure is more closely related to what stage of the task cycle the team is at when the AI fails. However, similar to situational awareness, the major change in our participants' perceptions occurred at a similar point. Our participants clearly had distinctive falloff points in failure performance during the action selection phase into the action implementation phase for both contexts, just before the upturns in perceived risk level by a phase. This indicates that decreased perceived failure performance leads to rising risk levels in a human-AI team task. This has important design implications, as there are various ways to simulate and measure a human's failure performance that could be used to predict when risk levels are about to increase to an unsafe level [39].

It is important to note that the failure performance and perceived risk levels of the participants saw the biggest changes as the agent entered the action selection stage, which seems to indicate something unique about this stage in the action cycle. Human-AI interaction research has shown that the action selection phase is subject to constant changes exhibited by the human with whom the AI interacts [40], and this variability may explain why these variables were most affected during this stage. This is because in teams, roles are highly interdependent, and one teammate's actions can immediately impact their teammate's considerations and decisions and even the team's overall goals.

VI. LIMITATIONS AND FUTURE RESEARCH

Previous research on the consequences of higher AI autonomy has produced conflicting results on how much AI autonomy affects teammate situational awareness and whether or not declines in that awareness truly impact human failure performance [15]–[17]. This study shows that discrepancy may be, at least partially, attributed to context. The AI in these experiments failed by missing something, as opposed to making a wrong action, and such ”false positives” should also be explored. While the participants in this study were all professionals with ample experience in their field, the study is limited in its sample size. Future research on whether the perceived risk levels shown in this study are consistent in larger sample sizes for risk type would allow AI designers to use perceived risk to triangulate AI autonomy levels between situational awareness and failure performance. Additionally, the sample utilized in this study focused primarily on professional experience and relied upon snowball sampling, resulting in a predominately white, Western perspective. Efforts to recruit and examine the perceptions of a more diverse sample should be considered in future research.

VII. CONCLUSION

As AI agents become more autonomous and assume unique roles on human-AI teams, researchers are tasked with understanding the risks and implications of their heightened autonomy. One such dangerous risk centers on the declining ability of human teammates to detect and respond to AI failures as their position in the AI's decision loop decreases. In this study, we conducted twelve contextual inquiries with professionals in the cyber incident response and medical professions. The study then examined how their situational awareness, failure performance, and perceived risk levels were affected by an AI teammate's autonomy level *and* the context in which the human-AI team operates. The inquiries revealed that while a human's situational awareness is more closely tied to context, failure performance and perceived risk level are more closely tied to the AI agent's functional autonomy level. These results provide valuable insight into how AI designers can use human perceptions of situational awareness, failure performance, and contextual risk to determine optimal autonomy levels for AI teammates.

REFERENCES

[1] A. I. Hauptman, B. G. Schelble, N. J. McNeese, and K. C. Madathil, "Adapt and overcome: Perceptions of adaptive autonomous agents for human-ai teaming," *Computers in Human Behavior*, vol. 138, p. 107451, 2023.

[2] M. Demir, N. J. McNeese, and N. J. Cooke, "Team communication behaviors of the human-automation teaming," in *2016 IEEE international multi-disciplinary conference on cognitive methods in situation awareness and decision support (CogSIMA)*. IEEE, 2016, pp. 28–34.

[3] M. M. Cummings, "Man versus machine or man+ machine?" *IEEE Intelligent Systems*, vol. 29, no. 5, pp. 62–69, 2014.

[4] J. B. Lyons, K. Sycara, M. Lewis, and A. Capiola, "Human–autonomy teaming: Definitions, debates, and directions," *Frontiers in Psychology*, vol. 12, p. 589585, 2021.

[5] R. Zhang, N. J. McNeese, G. Freeman, and G. Musick, "" an ideal human" expectations of ai teammates in human-ai teaming," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–25, 2021.

[6] N. J. McNeese, M. Demir, N. J. Cooke, and C. Myers, "Teaming with a synthetic teammate: Insights into human-autonomy teaming," *Human factors*, vol. 60, no. 2, pp. 262–273, 2018.

[7] A. K. Goel and L. Polepeddi, "Jill watson: A virtual teaching assistant for online education," in *Learning engineering for online education*. Routledge, 2018, pp. 120–143.

[8] L. Chan, L. Hogaboam, and R. Cao, "Artificial intelligence in video games and esports," in *Applied Artificial Intelligence in Business: Concepts and Cases*. Springer, 2022, pp. 335–352.

[9] M. Noriega, "The application of artificial intelligence in police interrogations: An analysis addressing the proposed effect ai has on racial and gender bias, cooperation, and false confessions," *Futures*, vol. 117, p. 102510, 2020.

[10] J. Zerilli, U. Bhatt, and A. Weller, "How transparency modulates trust in artificial intelligence," *Patterns*, p. 100455, 2022.

[11] B. G. Schelble, C. Flathmann, N. J. McNeese, G. Freeman, and R. Mallick, "Let's think together! assessing shared mental models, performance, and trust in human-agent teams," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. GROUP, pp. 1–29, 2022.

[12] T. O'Neill, N. McNeese, A. Barron, and B. Schelble, "Human–autonomy teaming: A review and analysis of the empirical literature," *Human factors*, vol. 64, no. 5, pp. 904–938, 2022.

[13] M. I. Ganesh, "The ironies of autonomy," *Humanities and Social Sciences Communications*, vol. 7, no. 1, pp. 1–10, 2020.

[14] C. D. Wickens, H. Li, A. Santamaria, A. Sebok, and N. B. Sarter, "Stages and levels of automation: An integrated meta-analysis," in *Proceedings of the human factors and ergonomics society annual meeting*, vol. 54, no. 4. Sage Publications Sage CA: Los Angeles, CA, 2010, pp. 389–393.

[15] G. A. Jamieson and G. Skraaning, "The absence of degree of automation trade-offs in complex work settings," *Human factors*, vol. 62, no. 4, pp. 516–529, 2020.

[16] ——, "The harder they fall? a response to wickens et al.(2019) regarding the generalizability of lumberjack predictions to complex work settings," *Human factors*, vol. 62, no. 4, pp. 535–539, 2020.

[17] C. D. Wickens, L. Onnasch, A. Sebok, and D. Manzey, "Absence of doa effect but no proper test of the lumberjack effect: A reply to jamieson and skraaning (2019)," *Human factors*, vol. 62, no. 4, pp. 530–534, 2020.

[18] A. I. Hauptman and N. J. McNeese, "Overcoming the lumberjack effect through adaptive autonomy," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 66, no. 1. SAGE Publications Sage CA: Los Angeles, CA, 2022, pp. 1075–1079.

[19] R. Parasuraman, T. B. Sheridan, and C. D. Wickens, "A model for types and levels of human interaction with automation," *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, vol. 30, no. 3, pp. 286–297, 2000.

[20] F. Flemisch, M. Heesen, T. Hesse, J. Kelsch, A. Schieben, and J. Beller, "Towards a dynamic balance between humans and automation: authority, ability, responsibility and control in shared and cooperative control situations," *Cognition, Technology & Work*, vol. 14, pp. 3–18, 2012.

[21] T. B. Sheridan, "Human centered automation: oxymoron or common sense?" in *1995 IEEE International Conference on Systems, Man and Cybernetics. Intelligent Systems for the 21st Century*, vol. 1. IEEE, 1995, pp. 823–828.

[22] M. R. Endsley and D. B. Kaber, "Level of automation effects on performance, situation awareness and workload in a dynamic control task," *Ergonomics*, vol. 42, no. 3, pp. 462–492, 1999.

[23] M. R. Endsley, "Automation and situation awareness," in *Automation and human performance: Theory and applications*. CRC Press, 2018, pp. 163–181.

[24] ——, "Toward a theory of situation awareness in dynamic systems," *Human factors*, vol. 37, no. 1, pp. 32–64, 1995.

[25] L. Onnasch, C. D. Wickens, H. Li, and D. Manzey, "Human performance consequences of stages and levels of automation: An integrated meta-analysis," *Human factors*, vol. 56, no. 3, pp. 476–488, 2014.

[26] M. R. Endsley, D. J. Garland *et al.*, "Theoretical underpinnings of situation awareness: A critical review," *Situation awareness analysis and measurement*, vol. 1, no. 1, pp. 3–21, 2000.

[27] K. E. Schaefer, E. R. Straub, J. Y. Chen, J. Putney, and A. W. Evans III, "Communicating intent to develop shared situation awareness and engender trust in human-agent teams," *Cognitive Systems Research*, vol. 46, pp. 26–39, 2017.

[28] P. Damacharla, P. Dhakal, S. Stumbo, A. Y. Javaid, S. Ganapathy, D. A. Malek, D. C. Hodge, and V. Devabhaktuni, "Effects of voice-based synthetic assistant on performance of emergency care provider in training," *International Journal of Artificial Intelligence in Education*, vol. 29, pp. 122–143, 2019.

[29] A. Baker, Y. Perov, K. Middleton, J. Baxter, D. Mullarkey, D. Sangar, M. Butt, A. DoRosario, and S. Johri, "A comparison of artificial intelligence and human doctors for the purpose of triage and diagnosis," *Frontiers in artificial intelligence*, vol. 3, p. 543405, 2020.

[30] M. M. Nyre-Yu, "Determining system requirements for human-machine integration in cyber security incident response," Ph.D. dissertation, Purdue University, 2019.

[31] L. Bilge, Y. Han, and M. Dell'Amico, "Riskteller: Predicting the risk of cyber incidents," in *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, 2017, pp. 1299–1311.

[32] E. H. Kahan, J. D. Shin, M. E. Jansen, R. H. Parker, and R. Parikh, "Malpractice cases arising from telephone based telemedicine triage in ophthalmology," in *Seminars in Ophthalmology*, vol. 38, no. 3. Taylor & Francis, 2023, pp. 294–305.

[33] P. Bednar and C. E. Welch, "Contextual inquiry and socio-technical practice," *Kybernetes*, 2014.

[34] C. E. Wells, "Teaching teamwork in information systems," in *Challenges of information technology education in the 21st century*. IGI Global, 2002, pp. 1–24.

[35] M. Jayashree and S. C. Singhi, "Initial assessment and triage in er," *The Indian Journal of Pediatrics*, vol. 78, pp. 1100–1108, 2011.

[36] O. Chirkov, N. Tsipina, S. Slinchuk, and E. Vorobyev, "Automated expert support complex based on a machine learning semantic processor," in *Journal of Physics: Conference Series*, vol. 1691, no. 1. IOP Publishing, 2020, p. 012062.

[37] E. M. Romeo, "Quantitative research on critical thinking and predicting nursing students' nclex-rn performance," *Journal of Nursing Education*, vol. 49, no. 7, pp. 378–386, 2010.

[38] C. Flathmann, B. G. Schelble, N. J. McNeese, B. Knijnenburg, A. K. Gramopadhye, and K. Chalil Madathil, "The purposeful presentation of

ai teammates: Impacts on human acceptance and perception," *International Journal of Human–Computer Interaction*, pp. 1–18, 2023.

[39] S. M. Jones-Jang and Y. J. Park, "How do people react to ai failure? automation bias, algorithmic aversion, and perceived controllability," *Journal of Computer-Mediated Communication*, vol. 28, no. 1, p. zmac029, 2023.

[40] G. Andrade, G. Ramalho, H. Santana, and V. Corruble, "Challenge-sensitive action selection: an application to game balancing," in *IEEE/WIC/ACM International Conference on Intelligent Agent Technology*. IEEE, 2005, pp. 194–200.