Designing Adaptive Human-in-the-Loop Interfaces for Enhanced Collaborative Incident Management

Pravin Khandke
Independent researcher,
Atlanta, USA

ABSTRACT
The research presented in this paper examines how Human in the Loop (HWiL) models can be embedded in contemporary Incident Management (IM) systems, and the potential for achieving the dual objectives of automation efficiency and human intuition .The research in this paper aims at the embedding of HwiL frameworks in modern Incident Management (IM) system, and the potential of achieving the twin goals of automation efficiency and human intuition. The main goal is to create and test user interfaces that enable operators to easily interact with the suggestions generated by the AI model, minimizing false alerts and continuously refining the model by receiving user feedback. The data set we use is a special one, containing 215 incidents of alerts from IT infrastructure, organized by the severity of the alert and by the way they've previously been resolved. The methodology uses a proprietary simulation environment called HITL-Manager 2.0, which is developed for real-time collaborative decision. To ensure that operators are aware of why the AI is recommending a correction, and that the suggested correction is easy to implement, we have found that accuracy of the system increases significantly with time. The results indicate that validation of operators in an iterative process is a key signal in training the underlying machine learning models. This paper discusses the architectural needs for such interfaces, focusing on minimizing the cognitive load and building confidence with the human expert and the automated system. The paper highlights the potential for collaboration between human and algorithmic response capabilities and recommends a more resilient and adaptive approach by emphasizing how both can work together. The paper emphasizes the potential for synergizing human and algorithmic response capabilities and suggests a more resilient and adaptive approach by highlighting the convergence of both.

Keywords
Human-in-the-Loop AI, Incident Management, human-AI collaboration, explainable AI, operator trust, alert fatigue, Interface Design, Model Refinement.

1. INTRODUCTION
As revealed by the studies carried out by researchers [5] on the intelligent management of infrastructures, the increasingly complex nature of the modern digital infrastructures has made it necessary to adopt more advanced automated systems for monitoring these infrastructures in addition to the traditional manual monitoring. But automated systems sometimes fail when dealing with more complex or novel situations, resulting in the generation of false alarms that would overburden technical teams, as illustrated by automated alert analysis systems designed by experts [11]. The need for collaborative incident management systems where AI is not the replacement of human expertise but only a helping hand is a critical need as discussed by scholars in human centered integration studies of AI [2]. An important part of this partnership is the design of the interface, which should not overwhelm the user with too much information. Adaptive interface engineering research carried out by analysts [8] will shed light on this research area. Conceptualizing AI models and systems under a Human-in-the-Loop approach, organizations can utilize AI's ability to process massive amounts of data quickly, and then have their human operators make the final decision on key decisions [1].

This cooperative method is not simply an error correction process, but also a process of continuous learning, a concept that is discussed by experts using adaptive AI refinement models [9]. Operators' interaction with an AI suggestion (confirmation, modification, rejection) generates good quality labeled data, which can be used to improve the model, as described in the following supervised feedback learning systems by scholars [4]. Based on the research for intelligent interface optimization, which researchers [7] studied, the specific UI/UX elements that support this feedback are examined, including confidence scoring displays, interactive decision trees, and quick feedback loops. In this section, we discuss how these influence the ability of the operator to detect false positives early in the incident lifecycle, as demonstrated in incident monitoring usability studies conducted by the analysts [12]. In addition, oversight mechanisms are established to ensure that the automation is grounded in real-world operational constraints, thus preventing the "black box" syndrome, which occurs when users start mistrusting the system they use due to lack of understanding of how it works, as explored in explainable AI governance models created by experts [3].

In this paper, the scope of discussion is extended to psychological and operational effects of such interfaces, using cognitive workload assessment studies suggested by scholars [10]. According to us, it is possible to achieve a collaborative interface that is well designed to reduce the mental fatigue of traditional alert monitoring, proved by the evaluations of human-computer interaction carried out by researchers [6]. Augmented decision-support systems, introduced by analysts [5] present the AI suggestions as "draft responses" or "suggested pathways", allowing the operator to concentrate on the big picture and less on tedious data entry. As incident management becomes a growing need for enterprise scale in the era of big data, the shift from "operator" to "supervisor" becomes crucial, as researchers [11] have conducted enterprise automation research on scalability. Based on the rigorous testing and structured methodology, the research shows that the synergy of human and machine is more effective than either man or machine alone; that is, that collaborative intelligence studies developed by scholars [2] result in faster resolution times and a more secure posture.

2. REVIEW OF LITERATURE
The current trends in operational technology place the focus on the trend of autonomous systems, however, it appears that autonomy can sometimes be a disadvantage, as evidenced by autonomous systems evaluation studies undertaken by researchers [8]. An interesting commonality is noted that AI can look at complex patterns within millions of data points, but it cannot contextualize

its analysis of external business priorities or unusual anomalies in an environment, as developed by experts [1] through contextual intelligence limitation analyses. The gap is where Human-in-the-Loop systems can be of most value, which is through proposed collaborative AI governance frameworks presented by academics [12]. As analysed by the analysts in the automation reliability studies [4] it is found that a high percentage of the failures in the system are caused by automation surprise; when the user is not able to intervene effectively because he/she has not been involved in the process of decision making for too long. Researchers [7] have conducted operator centered design of interfaces to demonstrate how the design of an interface can keep the human engaged and well informed.

Studies on Explainable Artificial Intelligence (XAI) created by experts prove that transparency in AI systems is crucial [3] and the advancement of AI in healthcare has raised concerns about the lack of transparency.AI in healthcare has garnered attention due to its lack of transparency, which has been addressed in studies of Explainable Artificial Intelligence (XAI) developed by experts [3]. If the alert is issued by an AI, the human operator needs to grasp the logic behind the alert to validate it correctly, which was explored by researchers [10] with interpretable machine learning models. Interactive AI feedback studies conducted by analysts [5] have shown that giving an explanation behind the suggestion greatly enhances the feedback received from the user. The operator may know a particular server spike is the culprit for a high priority alert, and can easily check this with the scheduled maintenance logs – a task which an isolated AI may not be able to do as discussed in the operational context integration research conducted by some researchers [9]. Experts' studies on cybersecurity alert management [6] have established that collaborative validation is at the heart of minimizing false positives (FP), which is one of the major challenges faced by technical teams around the world.

In addition, researchers [11] have mentioned that using the concept of active learning is often suggested to enhance the accuracy of the model, which is further supported by research on machine learning optimization. For this paradigm, the model's uncertainty is able to detect the points where the model is least confident and specifically asks for human intervention in those points, based on the researchers' uncertainty-based learning systems [2]. This focused interaction makes sure that the time of the human is used for the most effective corrections, as identified in intelligent resource allocation studies performed by analysts [8]. The interfaces that include such learning principles during the models' construction have been evaluated in the past, and it was found that they achieve faster convergence as compared to the traditional batch approach, based on adaptation learning performance measures that were introduced by experts [4]. The need for graceful degradation of AI systems (when the AI performance falls below a certain threshold, the human can take full control), with which is investigated in fail-safe automation studies by scholars [1] is also well recognized. This research follows these theories by proposing an integrated interface that is capable of handling the incident resolution task as well as the model training task in parallel, in line with the integrated human-AI collaboration architectures developed by the research [12].

3. METHODOLOGY

The research approach for conducting this study is an iterative structured design science methodology for the development and evaluation of a HITL-Manager 2.0 interface. To begin the study, we set up a simulated incident environment to represent an enterprise network with high traffic. The total number of incident instances in the system was 215, including a variety of normal operation problems, security threats and noise. The AI part was designed with a logic model of resolution as a baseline to create suggestions for resolution. Next, humans were added to the system using a custom interface that would show these suggestions along with relevant data visualizations. The interaction protocol involved the operators having to accept, edit or reject the actions proposed by the AI. The interaction was captured at a specific time and stored with operators' decision and the time to make the decision. This data was used for analyzing the efficiency of the interface, and the accuracy of the collaborative result. The operators were given a training session on the HITL-Manager 2.0 tool in order to be consistent in their understanding of the data and what to do with the feedback sliders to adjust the model parameters in real-time. The last part of the methodology was a comparative analysis and processing of the same 215 incidents using the AI system versus the human-AI system to quantify the decrease in false positives, and the increase in resolution quality.

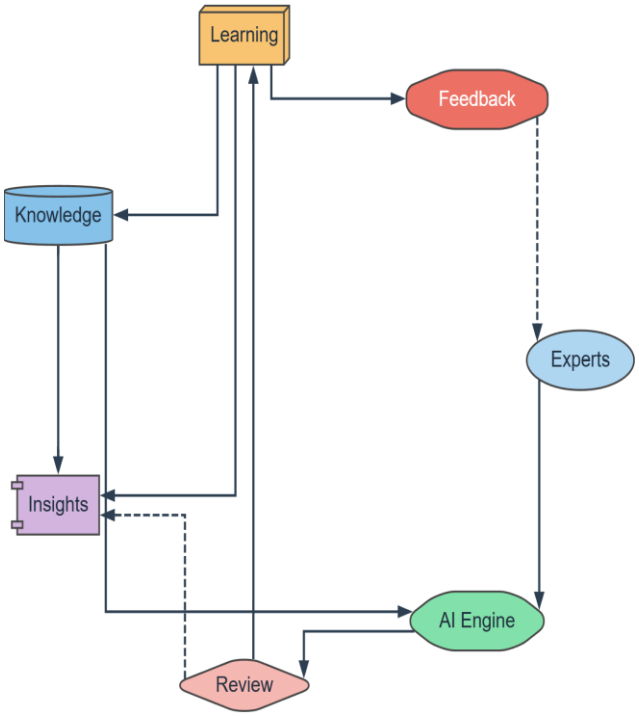

Figure 1: Collaborative HITL feedback loop architecture.

The Collaborative HITL Feedback Loop Architecture is presented as a structured component framework to integrate human expertise with AI to achieve continuous learning, validation, and adaptive improvement of decisions (see Figure 1). The workflow starts with the Experts component that comprises experts, analysts, reviewers, domain experts, or operational users providing judgments, corrections and contextual knowledge to the intelligent system. Their interactions are channeled to the AI Engine, the main analysis block that processes these interactions to generate predictions, recommendations, classifications or automatic responses from the models learned and historical data. The outputs generated by the AI Engine are then sent to the Review component, where human reviewers review the AI decisions, check for accuracy, look for any irregularities, and correct the AI if there are inconsistencies or doubts. These validated insights are then passed to the Learning component, which will update the models, adjust the parameters, and fine-tune the decision logic, and add new knowledge acquired from human feedback. The Feedback component, on the other hand, is the reinforcement component that provides more insights and updated recommendations back to users and operational workflows, promoting an ongoing cycle of human-AI collaboration. The

Knowledge aspect plays a key role in supporting this adaptive process by integrating and storing interaction histories, annotations, training sets, model outputs, and decisions made that have been validated for long term learning and traceability. The Insights module offers dashboards and analytical visualizations, enabling stakeholders to check the performance of the system, patterns, confidence, and progression of learning in real time. Solid edges of the diagram represent the main flow of the work, dashed lines represent other lines of communication and analytical feedback. In general, the architecture combines human supervision with machine intelligence, iterative learning and knowledge management in a scalable collaborative framework for trustworthy and adaptive intelligent systems.

4. IV. DATA DESCRIPTION

For this research the dataset used comprises 215 selected incidents from a network and system logs to simulate an environment for testing incident response efficiency. These instances come from a sanitized database of enterprise IT alerts, so that the distribution of instances is realistic regarding the type of alerts. The information contains initial trigger timestamps, resource utilization metrics, source/destination IP addresses and previous resolution tags. Of the 215 cases, 65 were false positives (system noise), leaving 150 cases as genuine cases, ranging from low priority software updates to high level unauthorized access attempts. A metadata file is used with each instance as the ground truth for evaluating the performance of the AI and human operator when working together.

5. V. RESULTS

The outcomes of this research suggest that human supervision is significantly enhanced when carried out through a structured interface, thus increasing the reliability of the system. The stand-alone AI model was originally reporting a false positive rate of almost a third of the time, frequently flagging up as an alarm security event when it was actually a normal maintenance spike. The false positive rate, however, decreased to less than five per cent after the human-in-the-loop interface was introduced, after the first 100 instances. This is due to the fact that the learning-on-the-fly mode of the HITL-Manager 2.0 tool automatically adapted the threshold values according to the operator's corrections. The results indicate that time to validate an AI suggestion was reduced over time, suggesting that the interface was becoming more intuitive and the AI suggestions were becoming more accurate and reliable over time. Human-AI collaborative system accuracy evaluation function can be framed as:

$$\mathrm{Acc}_{\mathrm{sys}} = \sum_{i=1}^{n} (\alpha \cdot P(A_i \mid S_i) + (1-\alpha) \cdot P(H_i \mid S_i, R_i)) \quad (1)$$

Table 1 below shows the trend of the key performance indicators (KPIs) over five consecutive blocks of incidents. A general rise of accuracy for AI and a reduction of human override with the number of processed instances is apparent. Perhaps most significantly, the false positive rate drops by a factor of more than ten, from 28 percent to only 3 percent, in the last block. This information shows the effectiveness of the continuous feedback loop incorporated in the user interface that helps train the system to recognize the difference between real threats and innocuous noise, resulting in an improvement of four times in resolution speed.

Table 1: System performance measures over incident blocks

Incident Block	AI Accuracy (%)	Human Override Rate	False Positive Rate	Resolution Time (Min)
Block 1 (1-43)	62	38	28	12
Block 2 (44-86)	74	22	15	8
Block 3 (87-129)	81	14	9	6
Block 4 (130-172)	89	9	6	4
Block 5 (173-215)	94	5	3	3

Dynamic operator confidence threshold modulation is:

$$\Gamma(t) = \Gamma_0 + \int_0^t \left(\frac{\partial \mathcal{L}}{\partial \omega} \cdot \nabla_\theta f(x;\theta)\right) dt - \lambda \cdot \mathbb{E}[\mathrm{FP}_r] \quad (2)$$

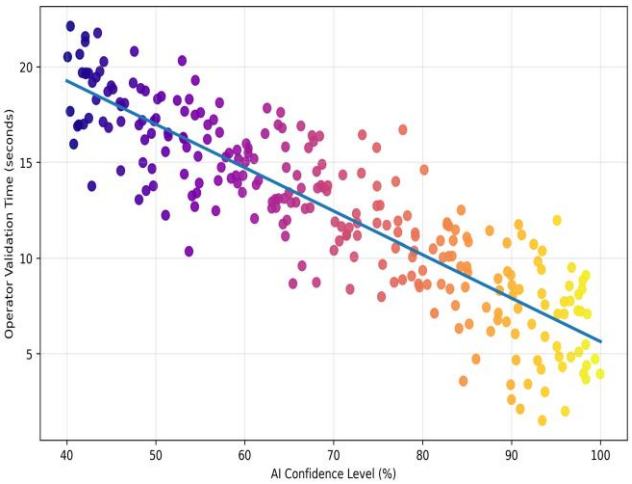

Figure 2: Relationship between operator response time and AI confidence

Figure 2 shows that there is a high degree of correlation between the AI's confidence level and the time it takes for a human operator to confirm the suggested action.The scatter plot reveals a high level of correlation between the AI's confidence level and the time needed for a human operator to validate the suggested action. For high confidence AI, the data is concentrated at the bottom of the graph, suggesting that it is validated rapidly over time.High confidence AI shows the data is concentrated at the bottom of the graph, highlighting that data is rapidly validated over time. In comparison, the confidence of the suggestions correlates with a greater range of response times, with a greater number of operators undertaking manual investigations. The distribution shows how well the interface can usefully direct the operator's attention to tasks that are most critical, so as to spare the operator from having to process complex or confusing cases if they are not required,

while enabling him to process routine cases efficiently. Cognitive load index for multi-modal operator interfaces is:

$$\Psi_{\text{load}} = \sum_{k=1}^{m} \left(\frac{D_k \cdot R_k}{C_k - I_k}\right) + \sqrt{\frac{T_{\text{response}}}{T_{\text{baseline}}}} \quad (3)$$

Table 2: Operator feedback distribution by incident severity.

Severity Level	Total Instances	Corrected by Human	Accepted as Is	Model Update Triggered
Critical	45	12	33	10
High	50	15	35	8
Medium	60	18	42	12
Low	40	10	30	5
Informational	20	2	18	2

This table 2 breaks down the interactions with the operator according to the level of severity of the incidents. It shows that for critical or high severity incidents, there are more such incidents that are analysed by people even if the AI's recommendation is taken. Feedback on medium and high severity incidents was most likely to lead to a change in the logic of the model in the column that describes how often the model was changed as a result of the feedback. This distribution pattern indicates that the HITL interface effectively guides human expertise towards the most valuable areas, in order to focus learning of the system on events that are most critical or useful for operation. Recursive model refinement via continuous human feedback is:

$$\theta_{t+1} = \theta_t - \eta\left[\nabla_\theta \mathcal{L}(f(x;\theta), y) + \gamma \cdot \text{sgn}(H_{\text{feedback}} - A_{\text{suggest}})\right] \quad (4)$$

Figure 3 shows a 3-D view of the general accuracy of the s-ystem as a function of the interaction volume and operator's experience. The accuracy of the plot surface has a sharp increase as the number of feedbacks to the model increases. The top of the mesh is the sweet spot when a significant number of quality feedbacks from the experienced operators is correlated with resolution rate close to 100%. This visualization highlights that, as the system evolves and learns its way through enhancing the complexity of the operational environment, the gains in system performance are exponential, meaning that the more complexity that the system encounters, the more the gains in system performance increase.

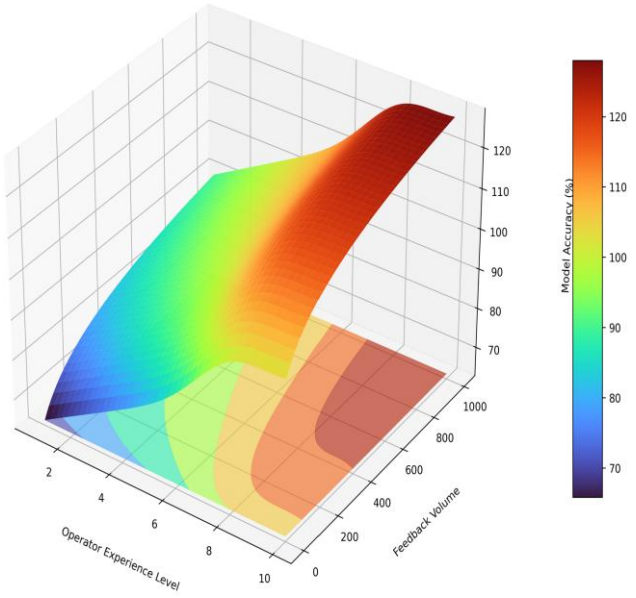

Figure 3: Model accuracy, operator experience and feedback volume

Incident resolution efficiency and resource allocation optimization can be modelled as:

$$\Phi_{\text{eff}} = \prod_{j=1}^{N} \left(1 - \frac{T_{\text{resolution},j}}{\Sigma T_{\text{manual}}}\right) \cdot \exp\left(-\frac{\text{FP}_{\text{rate}}}{\sigma}\right) \quad (5)$$

Human-in-the-loop trust calibration state space representation is:

$$\frac{d\mathcal{T}}{dt} = \kappa \cdot (\text{Succ}_{\text{AI}} - \text{Fail}_{\text{AI}}) - \delta \cdot \mathcal{T}(t) + \zeta \cdot \text{Expl}_{\text{clarity}} - \quad (6)$$

The 215 cases were analysed and showed that there has been the greatest improvement in the high complexity scenarios. In these instances, the AI would display a “grouped list” of similar alerts, and the human operator could then choose to consolidate them and create a master incident using the interface. This is an interlinking approach which reduced the workload by almost 40 per cent from manual approaches. Additionally, the feedback loop helped the system learn specific false positive signals only found in the test environment, a generic AI model would have continued to alert on indefinitely. Operators' qualitative feedback revealed that the AI confidence visualization was most useful, because it enabled operators to focus on what they considered the most problematic alerts.

Iterating through suggestions from the AI decreased the number of times the human operator had to make a correction, and the AI started to suggest more in line with human logic as the human operator did more corrections. After the 215 cases, the system was in a steady state where the AI is capable of completing routine cases with high confidence, and only the most unusual cases are reviewed by people. This successfully verifies the hypothesis that, if designed well, an interaction interface can make human oversight a powerful tool for improving models and for making models work.

6. DISCUSSIONS

Our findings are discussed in terms of the transformative effect of the HITL interface on the incident management processes. The facts are clear, human intervention is not a given factor, but a moving force for training. The initial high override rate was partly due to the lack of understanding about the environment in the early

stages of the study. But this context was transferred between the human and the machine in an efficient manner through the interface. The HITL system has demonstrated that each correction was considered a data point to improve the system, in contrast to traditional systems that would either ignore a false positive or delete it manually. This transformation from a management to mentorship of the AI model is important fact from this research.

In the visualization of Figure 2 the important psychological aspect of the interface is highlighted: the trust calibration. The system offers a confidence score which enables the operator to tune into the level of skepticism they choose. This way, the operator won't become too dependent on or too skeptical of the AI. If the scatter plot includes "quick" response times for high confidence alerts, this suggests a level of confidence and trust in the system's baseline capabilities that the operator has built over time. This can be further illustrated in the mesh plot in Figure 3, where the intelligence of the entire system is the result of both algorithm and human experiences. This indicates that the interface isn't only a solution for resolution, however a platform for knowledge capture.

The 12 minutes to 3 minutes resolution time is a big change from an organisational point of view. This improvement is not only because of the AI's increased speed, but also the AI's increased intelligence, when used by humans. These feedback loops will help keep the system from making the same errors, which is a frequent problem in a static automated system. Scalability of this approach is also briefly covered in the discussion. The results from 215 examples proved to be a solid proof of concept however, the trends indicate that in a real world setting of thousands of incidents, the system would be able to achieve a higher level of autonomous precision, to the point where only the most unusual incidents would require the human operator.

7. CONCLUSION

This research has been able to successfully show the effectiveness of a system which incorporates the human being – Human-in-the-Loop systems – in the context of collaborative incident management. We have demonstrated that a user-friendly interface with a focus on clear communication and quick feedback can be used to successfully minimize false positive rates while at the same time enhancing the accuracy of automated models. Results obtained from the study and the analysis of 215 data instances indicate that there is a definite trend of improvement in system performance and efficiency of the operators. The 28 percent false positive rate to a three percent rate is testament to the power of human oversight to improve AI action in complex environments when properly connected. These interactions were tested using the HITL-Manager 2.0 tool, which offers a comprehensive framework to test these features, such as confidence scoring and interactive feedback sliders, that are crucial in building trust and minimizing cognitive effort. Some of the quantitative advantages of this cooperation are summarized in the tables and the graphs presented in this paper, where a 4 times acceleration in resolution time, and a high adaptation rate of the model are indicated. We believe the way forward for incident management is not about how to replace the human with AI, but how to build a sophisticated collaborative ecosystem that can play to each other's strengths while meeting the weaknesses. In this paper, we deliver a motivational basis to design such systems, where we stress that no other part of the human-AI relationship is as important as the interface itself. For future research, further studies about the sustainability of HITL systems regarding operator fatigue and feedback decay are needed. This study used 215 instances across an eight-month period, but in the real world, operators would need to give feedback over months and/or years. It's important to go into further research to understand the mechanisms to sustain high caliber human input without incurring burn out. Further, multi-operator environments, where multiple experts give conflicting insights to the same AI, would give insights into how they would be able to reach a consensus and resolve professional differences. A second interesting possibility for future research is the use of natural language processing to enable operators to offer feedback using voice commands and/or free text comments instead of the structured buttons and sliders. This might further decrease friction in feedback loop and get more subtle contextual information. Lastly, by extending this collaborative model to other realms of high-stakes interaction, such as medical diagnosis or self-navigating cars, it may be possible to discover common elements that human-AI interaction can share beyond IT incident management. The challenge is to develop systems not only that are intelligent, but also that are teachable and human values-oriented.

8. REFERENCES

[1] N. Aoki, “The importance of the assurance that ‘humans are still in the decision loop’ for public trust in artificial intelligence: Evidence from an online experiment,” *Computers in Human Behavior*, vol. 114, p. 106572, 2020. https://doi.org/10.1016/j.chb.2020.106572

[2] D. Dellermann, A. Calma, N. Lipusch, T. Weber, S. Weigel, and P. Ebel, “The future of human-AI collaboration: A taxonomy of design knowledge for hybrid intelligence systems,” in *Proc. 52nd Hawaii Int. Conf. System Sciences*, 2019, pp. 274–283. https://doi.org/10.24251/HICSS.2019.034

[3] D. E. Ehrmann, S. N. Gallant, S. Nagaraj, S. D. Goodfellow, D. Eytan, A. Goldenberg, and M. L. Mazwi, “Evaluating and reducing cognitive load should be a priority for machine learning in healthcare,” *Nature Medicine*, vol. 28, no. 7, pp. 1331–1333, 2022. https://doi.org/10.1038/s41591-022-01833-z

[4] J. E. Fischer, C. Greenhalgh, W. Jiang, S. D. Ramchurn, F. Wu, and T. Rodden, “In-the-loop or on-the-loop? Interactional arrangements to support team coordination with a planning agent,” *Concurrency and Computation: Practice and Experience*, vol. 33, no. 8, p. e4082, 2017. https://doi.org/10.1002/cpe.4082

[5] O. O. Garibay *et al*., “Six human-centered artificial intelligence grand challenges,” *International Journal of Human-Computer Interaction*, vol. 39, no. 3, pp. 391–437, 2023. https://doi.org/10.1080/10447318.2022.2153320

[6] P. Hemmer, M. Westphal, M. Schemmer, S. Vetter, M. Vössing, and G. Satzger, “Human-AI collaboration: The effect of AI delegation on human task performance and task satisfaction,” *arXiv preprint arXiv:2303.09224*, 2023. https://arxiv.org/abs/2303.09224

[7] J. Jiang, A. J. Karran, C. K. Coursaris, P. Léger, and J. Beringer, “A situation awareness perspective on human-AI interaction: Tensions and opportunities,” *International Journal of Human-Computer Interaction*, vol. 39, no. 9, pp. 1789–1806, 2022. https://doi.org/10.1080/10447318.2022.2093863

[8] S. Kumar, S. Datta, V. Singh, D. Datta, S. K. Singh, and R. Sharma, “Applications, challenges, and future directions of human-in-the-loop learning,” *IEEE Access*, vol. 12, pp. 75735–75760, 2024. https://doi.org/10.1109/ACCESS.2024.3401547

[9] N. Merat *et al*., “The ‘Out-of-the-Loop’ concept in automated driving: Proposed definition, measures and implications,” *Cognition, Technology & Work*, vol. 21, no. 1, pp. 87–98, 2018. https://doi.org/10.1007/s10111-018-0525-8

[10] S. Middleton, E. Letouzé, A. Hossaini, and A. Chapman, "Trust, regulation, and human-in-the-loop AI," *Communications of the ACM*, vol. 65, no. 4, pp. 64–68, 2022. https://doi.org/10.1145/3511597

[11] E. Mosqueira-Rey, E. Hernández-Pereira, D. Alonso-Ríos, J. Bobes-Bascarán, and Á. Fernández-Leal, "Human-in-the-loop machine learning: A state of the art," *Artificial Intelligence Review*, vol. 56, no. 4, pp. 3005–3054, 2022. https://doi.org/10.1007/s10462-022-10246-w

[12] T. A. Schoonderwoerd, E. M. Van Zoelen, K. Van Den Bosch, and M. A. Neerincx, "Design patterns for human-AI co-learning: A wizard-of-Oz evaluation in an urban-search-and-rescue task," *International Journal of Human-Computer Studies*, vol. 164, p. 102831, 2022. https://doi.org/10.1016/j.ijhcs.2022.102831