Towards a Framework for Complementarity in Human-AI Collaboration

Anonymous Author(s)

Artificial intelligence (AI) can improve human decision-making in many application areas. Teaming between humans and AI may even generate complementary team performance (CTP), i.e., a level of performance beyond the one that can be reached by either one individually. Interestingly enough, CTP was only rarely achieved in previous research. We hypothesize that this is partially due to a missing holistic theoretical foundation that allows to systematically evaluate the factors influencing CTP. Therefore, we propose a theoretical framework based on two fundamental prerequisites for CTP—the existence of complementarity potential between humans and AI and an effective integration of both. Subsequently, we illustrate the utility of our framework. Our work provides researchers with a theoretical foundation of complementarity for human-AI collaboration.

CCS CONCEPTS • Human-centered computing → Empirical studies in HCI; • Computing methodologies → Artificial intelligence

Additional Keywords and Phrases: Human-AI Complementarity, Human-AI Collaboration, Complementary Team Performance, Human-AI Teams

1 INTRODUCTION

The increasing capabilities of Artificial Intelligence (AI) have paved the way for supporting human decision-making across a broad range of application domains. Examples include decision support for humans in the medical [1], legal [2], financial [3], and industrial domain [4].

The rise of AI often leads to a situation where an AI would theoretically be able to automate a particular task, such as diagnosing cancer. While automation could certainly be a goal in many areas, in others, complementing the human seems to be a more promising way forward, i.e., building a human-AI team [5]. In the example of cancer diagnosis, both humans and AI would be able to perform the overall task, but may have different performance on individual task instances, i.e., classifying a CT image. In the human-AI team, both the human and the AI will perform the cancer diagnosis and may be better at different instances of the task, complementing each other at the task instance level. While, for example, AI can detect patterns in vast large amounts of data that are difficult to discover for humans, humans excel in causal interpretation and intuition, which is necessary to interpret these patterns [6]. This complementarity potential has led researchers to investigate how the individual capabilities of humans and AI can be leveraged to achieve superior team performance compared to either one conducting the decision task independently. Such an outcome can be defined as complementary team performance (CTP) [7].

The majority of studies in the field merely demonstrate that humans teaming with AI may achieve higher team performance than conducting the decision task alone. Still, this performance is often inferior to the AI performance if it had performed the task alone [7], [8]. This leaves the question unanswered why CTP—exceeding this AI performance—could not have been accomplished.

In this context, the research community lacks a clear conceptual understanding of CTP, which prevents its rigorous evaluation. In this article, we therefore develop a theoretical framework to formalize complementarity. Our theoretical framework highlights the core influential factors of CTP by expanding existing frameworks [9]–[11]. Our work prepares the ground for the rigorous and fruitful development of design knowledge and artifacts for human-AI complementarity.

2 RELATED WORK

Complementarity between humans and AI is discussed in the context as part of several closely related paradigms.

First, Intelligence Augmentation focuses on elevating human capabilities, intelligence, and performance by having humans and AI each do what they do best. The idea is to design AI as systems that work with humans to enhance their performance compared to them conducting a task without any support [12].

Second, Human-Machine Symbiosis is a paradigm that envisions deepening the collaborative connection between humans and AI. It is based on the notion of a symbiotic relationship between humans and AI, which implies considering both as a common system rather than two separate entities with the goal to become more effective in comparison to both working separately [13]. This leads to overcoming human restrictions by extending their abilities and reducing the time needed to solve problems [14]. Another aspect further emphasizing the symbiotic nature of humans and AI is the focus on human-like communication and interaction between both team members. Therefore, the machine should be able to understand verbal and non-verbal communication to exchange information with the human [15].

Third, Hybrid Intelligence is an emerging paradigm that pursues the idea of combining human and artificial team members in the form of a socio-technical ensemble to resolve the current drawbacks of AI. We refer to the work of Dellermann et al., who defines Hybrid Intelligence as "the ability to achieve complex goals by combining human and AI, thereby reaching superior results to those each of them could have accomplished separately, and continuously improve by learning from each other" [16, p. 640].

Human-Machine Symbiosis and Hybrid Intelligence share our view that humans and AI should complement each other to achieve superior results. However, existing works under both labels neither provide theoretical foundations nor classify sources of complementarity potential and integration mechanisms.

The only work of which we are aware that contains a theoretical view of human-AI complementarity are the articles by Rastogi et al. [11] and Donahue et al. [10]. Rastogi et al. [11]develop a taxonomy characterizing differences between human and AI decision-making. Furthermore, the authors formalize one particular integration mechanisms in detail, the technical integration of human and AI decisions. However, they neither provide a formalization of those differences, nor a classification of the integration mechanisms. Donahue et al. [10] also focus on a single integration mechanism and formalize an integration algorithm that assigns weights to human and AI predictions. In summary, to our knowledge, there is no work that holistically formalizes the complementarity of humans and AI and provides classifications of sources and integration mechanisms. Moreover, both do not empirically evaluate their theoretical framework.

3 HUMAN-AI COMPLEMENTARITY

In this section, we introduce our proposed formalization of human-AI complementarity. Research has discussed multiple perspectives through which humans and AI models can complement each other [10], [16]. In this work, we focus on the performance perspective. In particular, we consider tasks that can be conducted by humans and AI models independently.

Let us consider a prediction task $T = \{x_i, y_i\}_i^N$ as a set of N instances $x_i \in X$ with a corresponding ground truth label $y_i \in Y$. The ground truth may not be known at the point of decision-making. However, it can be determined and revealed at a later point in time. Both a human decision-maker and an AI model are capable of independently inferring a prediction

$\hat{y}_i^H$ and $\hat{y}_i^{AI}$ for a given instance x_i. Additionally, let us consider some loss function l with its loss bounded in R^+. A loss function determines the error between a single prediction of human or an AI model and the corresponding ground truth label. In this context, it can be understood as a generic measure of task performance. In our formalization, it can encompass both classification as well as regression tasks. For a given prediction task, we denote the instance-specific human loss as l_H and the average loss over all available instances as $L_H = \frac{1}{N}\sum_{i=1}^{N} l_H(\hat{y}_i^H, y_i)$. Moreover, we denote the instance-specific AI loss as l_{AI} and the average loss considering all available instances as $L_{AI} = \frac{1}{N}\sum_{i=1}^{N} l_{AI}(\hat{y}_i^{AI}, y_i)$. For both the human and the AI model, we assume that their decisions are made independently.

In addition, we assume the existence of an integration mechanism $I(\hat{y}_i^H, \hat{y}_i^{AI})$, representing any way of collaboration between the human and the AI model. Collaboration between both team members results in an integrated decision which we denote as $\hat{y}_i^I$. This decision also incurs instance-specific loss l_I. Similarly, we define the average loss of the integrated decision as $L_I = \frac{1}{N}\sum_{i=1}^{N} l_I(\hat{y}_i^I, y_i)$.

Complementary team performance (CTP) results, once the average loss of the integrated decision is lower than both the individual average losses of the human and the AI model [7]:

$$CTP = \begin{cases} 1, & L_I < \min(L_H, L_{AI}), \\ 0, & otherwise. \end{cases}$$

In addition to this binary outcome, we interpret the difference between the loss of the best individual team member and the team performance as the amount of realized complementarity potential $CP_{realized} = max(0, min(L_H, L_{AI}) - L_I)$.

This measure forms the basis for developing a deeper understanding of the factors that can lead to CTP in human-AI collaboration. Specifically, we argue that $CP_{realized}$ is essentially composed of two components–the inherent complementarity potential ($CP_{inherent}$) between the human and the AI model, and the collaborative knowledge (CK) that can emerge through the collaboration itself.

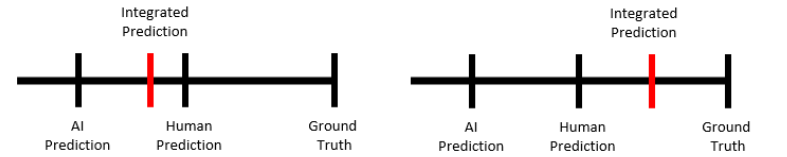
The first component can be understood as theoretically existing unique knowledge that the human and the AI model possess in relation to each other. From the perspective of the human, we denote it as unique human knowledge (UHK). It refers to the instances where the human achieves a lower loss compared to the AI model as the human can contribute unique knowledge on these instances. It can be defined it as the sum of the differences between the instance-specific AI model losses and the instance-specific human losses: $UHK = \sum_{i=1}^{N} max(0, l_{AI}^{(i)} - l_H^{(i)})$. From the perspective of the AI model, we denote it as unique AI knowledge ($UAIK$). It refers to the instances where the AI model achieves a lower loss compared to the human as the AI model can contribute unique knowledge on these instances. It can be defined as the sum of the difference between the instance-specific human losses and the instance-specific AI model losses: $UAIK = \sum_{i=1}^{N} max(0, l_H^{(i)} - l_{AI}^{(i)})$. From the perspective of the team member with the lower average individual loss, the other team members' unique knowledge then constitutes the theoretically existing unique knowledge that can materialize in improved team performance through human-AI collaboration. Thus, the inherent complementarity potential can be defined as

$$CP_{inherent} = \begin{cases} UHK, & L_{AI} < L_H, \\ UAIK, & L_H < L_{AI}. \end{cases}$$

In practice, it is unlikely that the integrated decision, resulting from human-AI collaboration, will always fully exploit $CP_{inherent}$. It represents an upper boundary of the first component of the realized complementarity potential $CP_{realized}$. However, in practice we are interested in knowing the amount that has been exploited by the integrated decision. We denote this as the inherent realized complementarity potential $CP_{inherent\_realized}$. To determine this amount, we must distinguish which of the two team members has the lower average individual loss.

When we consider $L_{AI} < L_H$, any integrated decision with $l_{AI} > l_I > l_H$ means that unique human knowledge is present, however, not fully exploited. In this case, $CP_{inherent\_realized}$ is the difference between the instance-specific loss of the AI model l_{AI} and the instance-specific loss of the integrated decision l_I. Figure 1a exemplifies this scenario. It

displays the absolute difference of the instance-specific loss of each team member individually and the integrated decision with respect to the ground truth for one particular instance. Moreover, in case the instance-specific loss of the integrated decision l_I even falls below the instance-specific loss of the human l_H ($l_{AI} > l_H > l_I$), $CP_{inherent\_realized}$ can only be the difference between the instance-specific loss of the AI model l_{AI} and the instance-specific loss of the human l_H (We will introduce the meaning of the difference between the instance-specific loss of the integrated decision l_I and the instance-specific loss of the human l_H when we elaborate the concept of collaborative knowledge). We visualize this scenario in Figure 1b.

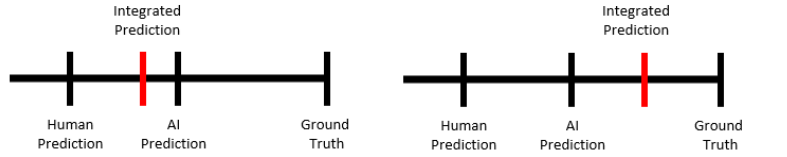


1a: Integrated prediction inside of AI and human prediction

1b: Integrated prediction outside of AI and human prediction

Figure 1: Human-AI collaboration combinatorics

Now, let us consider the case, $L_H < L_{AI}$. Any integrated decision with $l_H > l_I > l_{AI}$ can be interpreted that unique AI knowledge is present, however, not fully exploited. In this case, $CP_{inherent\_realized}$ is the difference between the instance-specific loss of the human l_H and the instance-specific loss of the integrated decision l_I. Figure 2a exemplifies this scenario. Again, in case the instance-specific loss of the integrated decision l_I even falls below the instance-specific loss of the AI model l_{AI} ($l_H > l_{AI} > l_I$), $CP_{inherent\_realized}$ can only be the difference between the instance-specific loss of the human l_H and the instance-specific loss of the AI model l_{AI}. We visualize this scenario in Figure 2b.



2a: Integrated prediction Inside of AI and human prediction

2b: Integrated prediction outside of AI and human prediction

Figure 2: Human-AI collaboration combinatorics

Finally, we can summarize each component of the inherent realized complementarity potential in the following formula:

$$CP_{inherent\_realized} = \sum_{i=1}^{N} \begin{cases} l_{AI}^{(i)} - l_I^{(i)}, L_{AI} < L_H \text{ and } l_{AI} > l_I > l_H \\ l_{AI}^{(i)} - l_H^{(i)}, L_{AI} < L_H \text{ and } l_{AI} > l_H > l_I \\ l_H^{(i)} - l_I^{(i)}, L_H < L_{AI} \text{ and } l_H > l_I > l_{AI} \\ l_H^{(i)} - l_{AI}^{(i)}, L_H < L_{AI} \text{ and } l_H > l_{AI} > l_I \end{cases}$$

The second component contributing to the realized complementarity potential $CP_{realized}$ is collaborative knowledge (CK). Collaborative knowledge refers to the idea that collaboration between the human and the AI model can result in an integrated decision l_I that improves for a specific instance beyond the existing unique knowledge in the team or that

deteriorates it even though one team member would have been able to provide a better decision individually. Therefore, we can distinguish positive collaborative knowledge (PCK) and negative collaborative knowledge (NCK) in the following way: $CK = PCK - NCK$.

Let us first focus on the positive collaborative knowledge. For a better understanding, let us again consider the scenario displayed in Figure 1b with $L_{AI} < L_H$. In case the instance-specific loss of the integrated decision l_I falls below the lower individual instance-specific loss of the human l_H ($l_{AI} > l_H > l_I$) or the AI model ($l_H > l_{AI} > l_I$), the difference between the lower individual instance-specific loss of either the human l_H or the AI model l_{AI} and the integrated decision l_I refers to positive collaborative knowledge as this improvement can only be driven by an act of collaboration and not by inherently present unique knowledge of one team member. The same phenomenon applies for the scenario displayed in Figure 2b with $L_H < L_{AI}$. In case the instance-specific loss of the integrated decision l_I falls below the individual instance-specific loss of the AI model l_{AI} ($l_H > l_{AI} > l_I$) or the human l_H ($l_{AI} > l_H > l_I$), the difference between the lower individual instance-specific loss of either the human l_H or the AI model l_{AI} and the integrated decision l_I refers to positive collaborative knowledge. We summarize positive collaborative knowledge in the following formula:

$$PCK = \sum_{i=1}^{N} \max\left(0, \min\left(l_H^{(i)}, l_{AI}^{(i)}\right) - l_I^{(i)}\right).$$

Lastly, we consider the scenarios in which negative collaborative knowledge (NCK) can occur. When we consider the scenario with $L_{AI} < L_H$. As in this scenario the AI model on average outperforms the human, we incur negative collaborative knowledge (NCK) once the instance-specific loss of the integrated decision l_I is larger than the instance-specific loss of the AI model l_{AI} ($l_I > l_{AI}$) independent of the instance-specific loss of the human l_H. In this case, negative collaborative knowledge (NCK) is the difference between the instance-specific loss of the integrated decision l_I and the instance-specific loss of the AI model l_{AI}.

Similarly, in the scenario with $L_H < L_{AI}$ the human on average outperforms the AI model. Thus, we incur negative collaborative knowledge (NCK), once the instance-specific loss of the integrated decision l_I is larger than the instance-specific loss of the human l_H ($l_I > l_H$) independent of the instance-specific loss of the AI model l_{AI}. In this case, negative collaborative knowledge (NCK) is the difference between the instance-specific loss of the integrated decision l_I and the instance-specific loss of the human l_H. We summarize negative collaborative knowledge in the following formula:

$$NCK = \sum_{i=1}^{N} \begin{cases} l_I^{(i)} - l_{AI}^{(i)}, L_{AI} < L_H \text{ and } l_I > l_{AI} \\ l_I^{(i)} - l_H^{(i)}, L_H < L_{AI} \text{ and } l_I > l_H \end{cases}$$

We are now able to better understand how the realized complementarity potential $CP_{realized}$ is composed of: $CP_{realized} = CP_{inherent\_realized} + PCK - NCK$. With this formalization, we are able to understand the innerworkings between the human and the AI model in the decision-making setting in greater detail.

4 ILLUSTRATION

In the previous Section, we introduced our formalization of human-AI complementarity. In the following, we discuss two scenarios to illustrate the benefits of our formalization.

We consider a task in which one human is assisted by an AI. In our two scenarios the AI is on average better than the human. To be more specific, on a test set with several task instances, e.g., 100 task instances, the AI on average performs better than the human. As a consequence, the inherent complementarity potential in our illustrative task depends on the unique human knowledge. For our scenarios, we choose an arbitrary regression task which could be for example a car price prediction task. We assume that the task has a clear ground truth. We use a sequential task setup: First the human makes an initial decision without receiving any AI assistance. Subsequently, the human receives an AI suggestion and is asked to

make a potentially revised second decision. This sequential task setup is necessary to measure our granular human-AI complementarity concepts, e.g., unique human knowledge. For both scenarios, we focus on a single task instance, i.e., one prediction of a car price. We keep the ground truth as well as the initial human decision and the AI suggestion constant and vary the final human decision and thereby introduce several sub-scenarios. By doing so, we keep the complementarity potential constant for all sub scenarios and highlight different results depending on the human-AI integration.

Scenario 1. In the first scenario (See Table 1), both the initial human decision and the AI decision are larger than the ground truth. The initial human decision (1,200.00 €) is closer than the AI (1,300.00 €) to the ground truth (1,100.00 €). We first analyze the complementarity potential present in the task instance. By staying with the initial human decision, the human would be 100.00 € better than the AI decision. This amount is the unique human knowledge present in the task instance.

In the following, we systematically vary the final human decision after receiving AI advice: In Scenario 1.1, the human decision after receiving AI advice is higher than the initial human decision and lower than the AI decision. In this task instance, the human can realize 50.00 € of the inherent complementarity potential.

In Scenario 1.2, the final human decision is smaller than the initial human decision and the AI prediction, thus being closer to the ground truth. This means we observe positive collaborative knowledge. Lastly, in Scenario 1.3, the final human prediction is worse than both the initial human and the AI decision. This means that the AI assistance essentially made the human worse than predicting alone. Besides those three sub scenarios, the final human prediction could also be lower than the ground truth. Similar to the collaboratively generated knowledge pattern, the human might decide finally based on new information to make a much lower prediction, e.g., a prediction of 1,050.00 € (Scenario 1.4). Our formalization also captures those edge cases. In this case, 50.00 € of positive collaborative knowledge would be present since our formulization takes the absolute distance from the ground truth into account.

Table 1 Scenario 1. Exemplary combinatorics for integrating human AI decisions. (CP: Complementarity Potential; PCK: positive collaborative knowledge; NCK: negative collaboration knowledge)

	Ground Truth	Initial human decision	AI	Human decision after receiving AI advice	Inherent CP	Realized inherent CP	PCK	NCK
Scenario 1.1	1,100.00 €	1,200.00 €	1,300.00 €	1,250.00 €	100.00 €	50.00 €		
Scenario 1.2	1,100.00 €	1,200.00 €	1,300.00 €	1.150,00 €	100.00 €		50.00 €	
Scenario 1.3	1,100.00 €	1,200.00 €	1,300.00 €	1,350.00 €	100.00 €			50.00 €
Scenario 1.4	1,100.00 €	1,200.00 €	1,300.00 €	1.050,00 €	100.00 €		50.00 €	

Scenario 2. In Scenario 2, the initial human decision is worse than the AI prediction which means it is further away from the ground truth. We first have a look at the complementarity potential of the task instance. By switching from the initial human decision to exactly the AI suggestion, the prediction would be 100.00 € closer to the ground truth (i.e., the unique AI knowledge in this task instance is 100.00 €). However, since the AI, per assumption, is better than the human on average, the complementarity potential of this task instance is 0.00 €.

Again, we are discussing several variations of the final human decision. In the first sub scenario (Scenario 2.1) the final human decision is improved. However, since we assume that the AI is on average better than the human, from the AI perspective no improvement happened. Therefore, also the realized complementarity potential is 0 €. In Scenario 2.2 the final human decision is better than the human and the AI decision, thus we again overserve positive collaborative

knowledge. Lastly, in Scenario 2.3 the final decision is worse than the initial human and the AI decision, thus negative collaborative knowledge is created.

Table 2: Scenario 2. Exemplary combinatorics for integrating human AI decisions. (CP: Complementarity Potential; PCK: positive collaborative knowledge; NCK: negative collaboration knowledge)

	Ground Truth	Initial human decision	AI	Human decision after receiving AI advice	Inherent CP	Realized inherent CP	PCK	NCK
Scenario 2.1	1,100.00 €	1,300.00 €	1,200.00 €	1,250.00 €	0.00 €	0.00 €		
Scenario 2.2	1,100.00 €	1,300.00 €	1,200.00 €	1.150,00 €	0.00 €		50.00 €	
Scenario 2.3	1,100.00 €	1,300.00 €	1,200.00 €	1,350.00 €	0.00 €			50.00 €

In an experiment, our formalization would be extended to multiple task instances and multiple individuals by averaging. Overall, these two toy examples have illustrated the utility of our formalization.

5 DISCUSSION

Our formalization allows us to generate insights on human-AI collaboration to make targeted design decisions.

Theory of Human-AI Complementarity. Our work advances the discourse on human-AI decision-making [7], [16] and in particular expands the theoretical groundwork. We provide a basis for future research in human-AI decision-making. In this context, we provide the research community not only with a theoretical formalization for human-AI decision-making but also with concrete measures that allow investigating the inner workings of human-AI decision-making from a nuanced perspective in behavioral experiments.

Design for Human-AI Complementarity. The most important implication of our research is the need to design for CTP, which is influenced by two factors–the complementarity potential and the human-AI interaction. Both can and should be designed. The inherent complementarity potential can be influenced by increasing the unique knowledge. From an AI perspective, this could, for example, be realized by designing "complementary" AIs that are particularly trained in those areas of the feature space where humans do not perform well [17]. Moreover, the integration mechanism needs to be consciously designed to maximize the collaboration knowledge.

Limitations. The sequential task setup necessary for our formalization has some disadvantages as it changes the task itself. Since conducting the same task initially alone before receiving AI advice, the human is already mentally prepared and might react differently than after directly receiving AI advice. The sequential task setup could induce an anchoring effect which prevents the human to more actively take the AI into account [18]. Moreover, sequentially conducted tasks with AI advice might not always be possible or desired in real-world settings. Therefore, the measurement should be seen as an approximation of real human behavior. Instead of having a sequential task setup, one alternative option could be to simulate a human model based on a data set of task instances solved by humans without AI advice. This simulation model could approximate the initial human decision within a non-sequential task setting. However, this approach is also an approximation of real human behavior. In other work, a latent construct has been derived to measure reliance behavior [19]. Future work should compare the approaches.

6 CONCLUSION

With an increasing number of tasks that can be automated, i.e., solved by AI alone, the focus will shift to the purposeful design of the collaboration between humans and AI as team members ("Human-AI teams"). The ultimate objective of these teams must be the achievement of complementary team performance (CTP) – with the team outperforming each individual team member. We hope that the conceptual foundations developed will provide fruitful ground for future research, and the empirical studies will illustrate the validity and potential of the human-AI complementarity paradigm.

REFERENCES

[1] L. Wu, L. Hitt, and B. Lou, "Data analytics, innovation, and firm productivity," *Manage Sci*, vol. 66, no. 5, pp. 2017–2039, 2020, doi: 10.1287/mnsc.2018.3281.

[2] K. Mallari, K. Inkpen, P. Johns, S. Tan, D. Ramesh, and E. Kamar, "Do I look like a criminal? Examining how race presentation impacts human judgement of recidivism," in *Proceedings of the 2020 Chi conference on human factors in computing systems*, 2020, pp. 1–13.

[3] D. Ahn, D. Lee, and K. Hosanagar, "Interpretable Deep Learning Approach to Churn Management," *Available at SSRN 3981160*, 2020.

[4] M. Stauder and N. Kühl, "AI for in-line vehicle sequence controlling: development and evaluation of an adaptive machine learning artifact to predict sequence deviations in a mixed-model production line," *Flex Serv Manuf J*, pp. 1–39, 2021.

[5] I. Seeber *et al.*, "Machines as teammates: A research agenda on AI in team collaboration," *Information and Management*, vol. 57, no. 2, p. 103174, 2020, doi: 10.1016/j.im.2019.103174.

[6] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, p. e253, 2017.

[7] G. Bansal *et al.*, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021, pp. 1–16.

[8] P. Hemmer, M. Schemmer, M. Vössing, and N. Kühl, "Human-AI Complementarity in Hybrid Intelligence Systems: A Structured Literature Review," *PACIS 2021 Proceedings*, 2021.

[9] A. Fügener, J. Grahl, A. Gupta, and W. Ketter, "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI," 2021.

[10] K. Donahue, A. Chouldechova, and K. Kenthapadi, "Human-algorithm collaboration: Achieving complementarity and avoiding unfairness," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1639–1656.

[11] C. Rastogi, L. Leqi, K. Holstein, and H. Heidari, "A Unifying Framework for Combining Complementary Strengths of Humans and ML toward Better Predictive Decision-Making," *arXiv preprint arXiv:2204.10806*, 2022.

[12] L. Zhou *et al.*, "Intelligence Augmentation: Towards Building Human-machine Symbiotic Relationship," *AIS Transactions on Human-Computer Interaction*, vol. 13, no. 2, pp. 243–264, 2021.

[13] J. C. R. Licklider, "Man-computer symbiosis," *IRE transactions on human factors in electronics*, no. 1, pp. 4–11, 1960.

[14] A. Gerber, P. Derckx, D. A. Döppner, and D. Schoder, "Conceptualization of the Human-Machine Symbiosis – A Literature Review," *Proceedings of the 53rd Hawaii International Conference on System Sciences*, vol. 3, pp. 289–298, 2020, doi: 10.24251/hicss.2020.036.

[15] J. C. Sanchez and J. C. Principe, "Prerequesites for symbiotic brain-machine interfaces," in *2009 IEEE International Conference on Systems, Man and Cybernetics*, 2009, pp. 1736–1741.

[16] D. Dellermann, P. Ebel, M. Söllner, and J. M. Leimeister, "Hybrid Intelligence," *Business and Information Systems Engineering*, vol. 61, no. 5, pp. 637–643, 2019, doi: 10.1007/s12599-019-00595-2.

[17] H. Mozannar and D. Sontag, "Consistent estimators for learning to defer to an expert," in *International Conference on Machine Learning*, 2020, pp. 7076–7087.

[18] Z. Buçinca, M. B. Malaya, and K. Z. Gajos, "To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making," *Proc ACM Hum Comput Interact*, vol. 5, no. CSCW1, pp. 1–21, 2021.

[19] H. Tejeda, A. Kumar, P. Smyth, and M. Steyvers, "AI-Assisted Decision-making: a Cognitive Modeling Approach to Infer Latent Reliance Strategies," *Comput Brain Behav*, Dec. 2022, doi: 10.1007/s42113-022-00157-y.