UHASSELT
KNOWLEDGE IN ACTION

Faculty of Business Economics

Master of Management

Master's thesis

Hybrid Intelligence. The strength and challenges of getting the human in the AI Loop, a literature study

MISS BOONYA CHIVAPONG

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :
Prof. dr. Benoit DEPAIRE

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

2019
2020

Faculty of Business Economics

Master of Management

Master's thesis

Hybrid Intelligence. The strength and challenges of getting the human in the AI Loop, a literature study

MISS BOONYA CHIVAPONG

Thesis presented in fulfillment of the requirements for the degree of Master of Management, specialization Business Process Management

SUPERVISOR :

Prof. dr. Benoit DEPAIRE

RESEARCH PAPER

Hybrid Intelligence: The Strengths and Challenges of Getting Humans in AI Loop

Miss Boonya Chivapong

Universiteit Hasselt

Missboonya.chivapong@student.uhasselt.be

Abstract

Hybrid Intelligence is the exploitation of Human Intelligence and Artificial Intelligence combination. The emergence of Hybrid Intelligence allows both entities to overcome their limitations and, at the same time, leverage each individual's strengths to achieve superior results that cannot be achieved by each of them in separation. However, there are some shortcomings in Human-Machine Teaming (HMT), as humans and AI are two dynamic entities that have distinct mental models, expertise, and abilities. Thus, developing Human-in-the-loop (HiL) for the future should take into consideration on how to bridge the gap between the two entities and how to put Human-in-the-loop AI into practice for real-world problems. The contribution of this paper is threefold: 1) to categorize and analyze the potential of state-of-the-art Human-in-the-loop AI in different application areas; 2) to categorize and analyze the challenges and limitations of the existing Human-in-the-loop systems; 3) to provide a guideline for future research on directions of Hybrid Intelligence.

Keywords: Human-in-the-loop (HiL), Human-Machine Teaming (HMT), Human-Computer Interaction (HCI), Machine Learning, Decision-making

COVID-19 Crisis Disclaimer

This master thesis was written during the COVID-19 crisis in 2020. This global health crisis might have had an impact on the (writing) process, the research activities and the research results that are at the basis of this thesis.

1. Introduction

1.1 Background

Nowadays, Artificial Intelligence has become an inevitable yet crucial part of our lives. Information and data are much easier and faster to access across the globe through mobile phones and electronic devices. For example, O2O (Online-to-Offline) technology, which connects the physical world to the digital world, has played an important role in our daily lives, especially for metropolitans. Therefore, the only way to make the best use of it is to explore and understand how AI works to adapt and enjoy living in a modern world with the company of AI, which, mostly, will make our lives more convenient and less burdensome in several aspects.

Despite its wide range of capabilities and consistent productivity, humans are still left doubting with some vague concepts of AI's ethics, security, fairness, and trust. Nowadays, machines can overcome difficulties in complex tasks without human intervention. Instead, their performance has developed over time by Machine Learning, which means the machines possess self-learning ability that allows them to train their algorithms by repeating the process until the error rate is lower than the expected threshold value.

In some fields, especially the ones that intimately related to the safety of life, for instance, medical diagnosis and autonomous vehicles, the importance of accuracy is crucially decisive since our lives would heavily depend on the decision that the machines have made.

Although machines outperform humans in tasks like synthesizing and processing a large amount of data, spotting weak features, predicting from hidden correlations and helping humans making decisions in the world of increasing complexity, humans, on the other hand, are better than machines in the aspects of decision-making under uncertain circumstances (Dellermann and Calma, 2018), high dexterity tasks (Lee, 2018) and many more. Therefore, the credibility of having humans co-perform the critical tasks, such as high-stake decision-making, will be the key to increase the confidence of the accuracy because humans can, at the same time, supervise and fix the errors or mistakes caused by machine failure. Moreover, involving humans in the decision-making process can also ensure the trustworthiness of the results as people would rather accept the final result provided by humans than that of AI, which leads to the development of Hybrid Intelligence.

In this paper, we start by discussing “What is Human-in-the-loop AI?”, then followed by “Why do we need Human-in-the-loop AI?” as to develop some basic understandings and gain some background knowledge of Hybrid Intelligence or the so-called Human-in-the-loop AI. After that, we discuss the strengths of getting humans into the AI loop, which are discussed further in several aspects. Ethics, Security, Trust, and Fairness is the perspective that crucially concerned by most people. The Credibility of Humans in Decision-Making Tasks, however, is the main reason why Human-in-the-loop AI is necessary. The most significant result of getting humans into the AI loop is The Improvement of Task Performance by involving Humans in Machine Learning, which will reflect in The Benefits of Co-Learning over time. However, The Mismatch between humans and AI is one of the biggest obstacles for performance optimization. Combining two different entities means that the pros and cons of each are as well integrated, which reflect in Human Errors and Machine Failures, which can also be caused by the imperfect Human-AI Interface and solved by Human Behavioral Model, but modeling human behaviors is still far from being perfect. In the end, The Future of Human-in-the-loop AI is provided in the three main questions based on how to successfully develop Hybrid Intelligence for the future.

1.2 Research Questions

This paper investigates the following problems:

Main Research Questions

RQ 1: Which are the potential and challenges of humans in the loop AI in comparison to the traditional AI?

Sub-Questions

RQ 1.1: What is Human-in-the-loop AI?

RQ 1.2: Does Human-in-the-loop AI perform better than each of them alone?

RQ 1.3: How can Human-in-the-loop AI improve the performance of the traditional AI?

RQ 1.4: Why does getting humans into AI loop become challenging and costly?

1.3 Methodological Approach

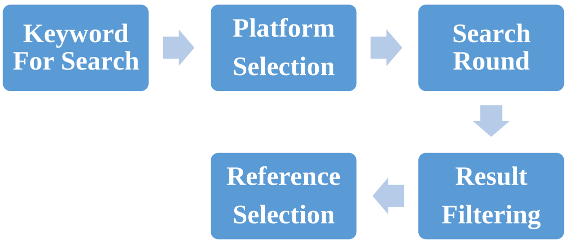

Figure 1: The approach of sample collection

In this research, the approach for sample collection was partly based on (Webster and Watson, 2002), and had been adapted to the strategies of (Ridley, 2012).

The literature search has been conducted from November 2019 to March 2020 with 17 different keyword combinations, such as human-in-the-loop, Hybrid Intelligence, Human-AI collaboration, Human-Machine Teaming, Human-Computer Interaction, Interactive Machine Learning, human helper, Human-Machine Interaction, human decision-making, human-centered, XAI, and others, from the field of HCI, HMT, XAI, and iML, accessing through 4 reliable online platforms: Google Scholar, IEEE Computer Society Digital Library, Springer, and arXiv. The samples are only collected from the free database or free-access journals on the websites listed, also from the physical books that I owned or borrowed from the university library. The sample collection has been conducted through 23 search rounds (from November 2019 to March 2020) to select the most relevant and up-to-date articles based on the topic and research questions. However, direct keyword combinations like "Hybrid Intelligence Strength" or "Hybrid Intelligence" did not come up with many related results. By using all the sets of keywords, 356 relevant results have been found collectively via 4 different platforms mentioned earlier. After primary filtering by reading the abstract and skimming through the whole article, 108 strongly related articles that mainly focus on Human-in-the-loop AI and its strengths or weaknesses were kept for the second-round filtering. For the second-filtering, going through all the studies thoroughly and expand further reading by following some conceptual in-text citations of each article helps to extract and select the most strongly relevant studies, which have been through peer-evaluation (however, only a few of useful articles selected have not been published officially). Finally, 57 most related information sources (3 books, 50 articles, and 4 websites) are applied to this literature review.

1.4 Structure

The main contribution of this study is to bring critical considerations from the field of Hybrid Intelligence. Consequently, this paper is structured as follows. Section 2 proposes a general idea of what Human-in-the-loop AI is, why we need Human-in-the-loop AI, and draws a comparison between Human Intelligence and Artificial Intelligence. Section 3 provides a holistic overview of the strengths of Human-in-the-loop AI. Section 4 addresses the challenges encountered by getting humans into the AI loop. Section 5 discusses the future direction of Human-in-the-loop AI. And section 6 summarizes the paper.

2. Why Human-in-the-loop?

To answer the question of "why do we need Human-in-the-loop AI?", it is necessary to develop a clear understanding of the concept of Hybrid Intelligence. But before that, it is required some background knowledge of Artificial Intelligence (AI).

The definition of Artificial Intelligence (AI) might somehow be ambiguous since no one knows how to correctly define the word *intelligence*. However, (Luger and Stubblefield,1998) suggested the definition of Artificial Intelligence (AI) as the branch of computer science that is concerned with the automation of intelligent behavior. Artificial Intelligence mainly focuses on problem-solving and also can be applied in various areas, namely Game Playing, Automated Reasoning and Theorem Proving, Expert Systems, Natural Language Understanding and Semantics, Modeling Human Performance, Planning and Robotics, Languages and Environment for AI, Machine Learning, Alternative Representations: Neural Nets and Genetic Algorithms and so on.

After acquiring some background knowledge of Artificial Intelligence, now it is time to look further into the concept of Hybrid Intelligence. Hybrid Intelligence (HI) is the ability to accomplish complex goals by combining human and artificial intelligence to collectively achieve superior results and continuously improve by learning from each other (Dellermann et al., 2019a). The core of Hybrid Intelligence is the combination of human and machine intelligence, expanding human intellect instead of replacing it. Hybrid Intelligence takes human expertise and intention into account when making meaningful decisions and taking appropriate actions conforming to ethical, legal, and societal values. The key design of Hybrid Intelligence is getting humans into the AI loop, so the main focus of this paper is Human-in-the-loop (HiL), which refers to human interaction in the AI loop.

Apart from the definition of Human-in-the-loop AI, another essential aspect of Human-in-the-loop AI is its applications. According to (Munir et al., 2013), there are three main categories of applications:

1) applications where the system is directly controlled by humans called supervisory control;

2) applications where the system passively monitors humans and takes appropriate actions (including doing nothing), it could be either open-loop or closed-loop system;

3) the combination of the prior and the latter called a hybrid system.

The taxonomy of Human-in-the-loop AI's applications based on the role of the human in the loop, whether the human takes the initiative to control, being monitored passively while the system takes action, or both. Thus, more details of the applications will be further discussed in the next section.

Why do we need Human-in-the-loop AI? There are two possible ways to answer this question. One possible way is to look at it from the perspective of subjective humans, most of the people are always aware, or in other words, afraid of the uncertainty of the unknown, which leads to the idea of merging themselves to the AI by the threat of possible job losses. On the other hand, from the perspective of AI development, AI has struggled with some limitations over the past decades, and humans may be of help.

From the past few years, AI has already boosted the efficiency of specific tasks in various industries or even proceed with the tasks faster and more precisely than human experts, yet, at the same time, it could be a threat to our job market as well.

From the perspective of subjective humans, accepting the fact that AI started to take away some of our jobs could make some of us insecure about our future. (Lee, 2018) has described the characteristics of jobs that would possibly be wiped out of the job market as asocial and optimization-based, such as insurance adjuster, personal tax preparer, customer service representative, radiologist, basic translator, and many others. (Zanzotto, 2019) believes that AI has to pay back to the society what it has "stolen" and portrayed Human-in-the-loop AI (HitAI) as a fairer approach that can compensate to the knowledge producers what is owed. Data pool is the most valuable treasure of AI in the process of Machine Learning, and we—skilled or unskilled workers are consciously, or sometimes, unaware of providing AI knowledge that it can benefit from without receiving payment in return. In contrast, the knowledge and data provided by humans would make AI become more powerful and eventually replace some of the repetitive jobs, which is why humans need to merge into AI to leverage and regain what belongs to us.

However, the reality might not be as extreme. (Lee, 2018) believes that the jobs that remain for humans are those required highly social interaction and those in an unstructured environment with high dexterity. This is because humans can learn from experience that does not require a large scale of samples to process compared to AI. Commonsense is always what makes humans different from AI, which reflects in humans' abilities in dealing with the unstructured environment while AI cannot, and that is the main reason why we, most of the time, need humans in the decision-making process.

As people who are always able to adapt to changes would never or hardly lose their jobs, or in other words, able to find a new type of jobs that AI would bring into the market, so integrating humans into the AI loop is an indirect way to prevent unpredictable unemployment in the future. Even though this psychological aspect is worth mentioning, yet it may not be significant enough to be the main focus of this paper.

On the other hand, there is the other, more widely accepted, way to explain why we need Human-in-the-loop AI. From the perspective of AI development, although there are various types of tasks that AI performs better than humans, still some tasks cannot be achieved by AI alone, which is why we need Human-in-the-loop AI to overcome these limitations.

Recently, Autonomous AI wave started washing over the global AI market, autonomous robotics applications began to enhance and disrupt the global job market in various industries contemporaneously. Speaking about Autonomous AI, most people would come up with autonomous vehicles, which believed to reduce road accidents and fatality. However, such a promising innovation unsurprisingly has its drawbacks. Despite its excellent performance, when facing tradeoff decisions like "to live or to die" or "whose life to sacrifice" (Lee, 2018), how can AI manage to make the wisest decision and on which criterion to be based? These questions thus far remained unanswered. According to the aforementioned cases, Artificial Intelligence can solve complex tasks better than humans in a shorter amount of time. But taking into account all the limitations leading to trust, security and ethical problems, that is where humans should intervene or take part.

Although AI performs better in the areas of complex problem-solving requiring a higher level of mathematics skill and hidden correlations, synthesizing and processing large amounts of data, and other consistent tasks (Lee, 2018), AI can still benefit from human assistance in cognitive tasks, for instance, map labeling (Klute et al., 2019), feature selection (Correia and Lecue, 2019), annotating arbitrary data (Hillen and Höfle, 2015), solving highly uncertain tasks, complex decision making and some other tasks that required *gut feeling*, such as reasoning, language understanding, complex decision-making in highly uncertain circumstances (Dellermann and Calma, 2018).

Most studies stated that integrating humans with machines will improve efficiency and outperform each of them alone. Like in the gaming area, in a chess tournament, there were three types of team combination—Human-Human, Machine-Machine, and Human-AI. The result is that the Human-AI combined team won and performed better than both humans and machines alone (Baraniuk, 2015). Consequently, we are convinced by such incredible stories of Human-AI collaboration enhance the performance of the standalone individual.

Besides, when the machines are left functioning alone, they occasionally make some mistakes or errors, which can directly affect user trust, thus getting humans in the AI loop will decrease the failures of the machine, and feedback from humans can help the machine improve and learn throughout its life cycle (see further in section 3), which is why leveraging Human Intelligence to augment Artificial Intelligence application—Humans and AI collaboration (or the so-called "Hybrid Intelligence" or "Human-in-the-loop AI") is a promising solution to this problem.

Although Human-in-the-loop AI has been developed to enhance the potential of the system, getting humans in the AI loop has its complexity because integrating humans into the framework means that the weaknesses of humans are also absorbed (more details in section 4).

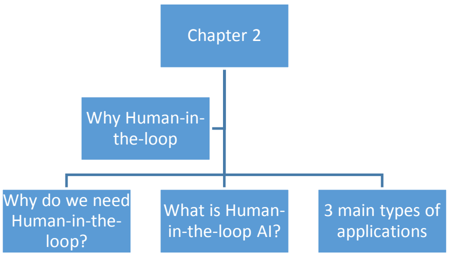
Chapter 2
Why Human-in-the-loop
Why do we need Human-in-the-loop?
What is Human-in-the-loop AI?
3 main types of applications

Figure 2: Main points of Chapter 2

In conclusion, this chapter has covered the questions of "What is Human-in-the-loop AI?" by providing common definitions of Hybrid Intelligence; "Why do we need Human-in-the-loop AI?" by listing weaknesses of traditional AI and explaining how getting humans into the AI loop can benefit the performance of the traditional AI; and "3 main types of applications" based on the role of the human in the loop.

3. The Strengths of Human-in-the-loop

In this section, we will discuss the potential of Human-in-the-loop AI from several aspects. Ethics, security, fairness, and trust (fairness in this case refers to the fairness of the final decision/result, which could be determined by the transparency of the decision-making process) are the most decisive factors in determining whether the results produced by Human-Machine Team (HMT) are reliable or not, also the algorithmic bias and human bias are as well taken into consideration. Moreover, the impact of getting humans into the AI loop is significant in various application areas. Furthermore, humans can reduce complexity in Machine Learning. In addition, by integrating humans into the framework, humans and machines can continuously develop and learn from each other to achieve superior results.

3.1 Ethics, Security, Fairness, and Trust

AI ethics is a set of values, principles, and techniques that employ widely accepted standards of right and wrong to guide moral conduct in the development and use of AI technologies. These values, principles, and techniques are intended both to motivate morally acceptable practices and to prescribe the basic duties and obligations necessary to produce ethical, fair, and safe AI applications (Leslie, 2019).

In this section, we mainly focus on the impact of biased systems that produced unfair results, the importance of security when encountering tradeoffs related closely to our lives and well-being, and how to develop human-in-the-loop AI towards a higher level of ethics, security, fairness, and trust.

In the age of AI-driven society, AI has become such a powerful tool that can make critical decisions in our lives. In the finance industry, Cindicator has developed a Hybrid Intelligence infrastructure by combining human financial analysts with machine learning models to help investors manage their assets more effectively in such a volatile economy and uncertainty (Cindicator, 2017). In the legal industry, China has implemented Artificial Intelligence on the jurisdiction and deployed its first digital court (Cyber court) in Hangzhou, which later has been launched into 12 provinces and regions (Deahl, 2017).

However, when encountering the situation where human life is at stake, how well AI can deal with a tradeoff is the most critical aspect to bring in concerns. For example, when facing a tendency of car accidents, the semi-autonomous vehicle can choose to protect the human driver at the expense of the nearest human operator's life, but the problem is "Is it an ethical/moral decision to make?".

The importance of implementing Human-in-the-loop AI for this problem is that when an accident happens or when the automated system malfunctions, the human would be the one to bear the brunt of the moral and legal responsibilities (Elish, 2019). Apart from that, Human-in-the-loop AI can ensure the security of semi-autonomous and autonomous vehicles by developing the driver model to further predict the driver behavior, which could lead to the measurement to handle infrequent events and variances in driving scenarios (Driggs-Campbell et al., 2015).

Therefore, AI systems must be built in ways that ensure that humans are always in ultimate control and responsibility for all that the AI system will do. This is particularly significant with regard to decisions that affect a person's life, quality of life, health, or reputation. All decisions and outcomes must remain the designated responsibility of humans. This is both to ensure that the decision is made carefully, but also to maintain the role of AI systems in supporting humans. The goal of designating specific responsibility is to maintain human control and increase personal investment in the product (Smith, 2020).

Most people believe that the decisions made by algorithms are more objective and fairer. However, the algorithms can also generate algorithmic bias, such as Gender bias and Racial/Ethnic bias that are commonly introduced to the system through the learning process. Thus, the biases inherited in the AI system can affect the fairness and the quality of the decisions.

How can algorithms generate biases? Algorithmic bias can be derived from different sources, for example, biased dataset input, unrepresentative dataset, error minimization attempt, and sensitive attribute (Pessach and Shmueli, 2020).

In the learning process, the models are fed with a massive amount of input data, and it is proven that there are historical biases included in the dataset since some piece of information is processed by humans. Therefore, when the machine learns from the biased data, it may be inherently biased, which means human bias can be transferred to the algorithms through the learning process.

This can be primarily solved by Bias Rating of AI Services (Srivastava and Rossi, 2019), which can differentiate the biasness of AI services into three levels—unbiased compensating (the system itself is unbiased and does not follow biases in the input data), data-sensitive biased (the system would follow the biases included in the dataset), and biased (the system itself is bias and can introduce biases even when processing unbiased dataset).

Understanding the level of the biasness in the system allows us to manage the set of information by reducing or removing biases from the input data, and with this human intervention, the quality of the service or the fairness of the result can be improved.

Taking an example of how human intervention can reduce algorithmic biases, Gender bias is one of the most common problems faced by many companies and industries, especially in some occupations that are more representative of one gender than the other. For instance, most of the engineers, technicians, and mechanics are presumably male, while most of the nurses are assumed to be female. Thus, the algorithm undoubtedly would over-represent one gender over the other based on the bias injected.

To reduce the biases in a dataset without affecting its accuracy, integrating humans in the decision pipeline is crucial since there is a tradeoff between fairness and accuracy. As we pursue a higher degree of fairness, we may compromise accuracy (Kleinberg et al., 2017).

To solve the problem of gender bias in recruitments, (Peng et al., 2019) found that human intervention can reduce biases to a certain level. By balancing gender slates, over-representing, and under-representing. However, these methods can mitigate bias in some professions, but not all.

It is almost impossible to get rid of all the biases, especially in those biased systems (the worst rating of all three). Therefore, the next question to answer is “How to make a strong Human-Machine Teaming (HMT) that conforms to ethics, security, fairness, and trust?”

Designing a model that conforms to ethics, security, fairness, and trust would require a fair amount of technical skills to develop an accountable system that is ethical, safe, and fair. Once the accountability of the system is achieved, the security of the human users is more guaranteed.

In order to gain user trust, only ensuring ethics, security, and fairness might not be enough. Transparency in decision-making processes also plays an important role in trust forming for both users and the human in the loop. As a teammate, the human must understand the logic behind every decision AI made so that he/she can take control when unexpected situations occur or even prevent harmful use and algorithmic failures by closely monitoring them (Chakraborti et al., 2017).

Once the trust is formed, it will reflect on the higher level of teamwork performance as Human-Machine Teams are strongest when humans can trust AI systems to behave as expected, safely, securely, and understandably.

Taking Recommender Systems for example, when such systems properly address the issues of Fairness, Accountability, Transparency, and Ethics, then the trust of the user in the system would just depend on the system's output (Pelta et al., 2020). User trust can develop over time when continuously receiving good recommendations from the recommender system. On the other hand, when consistently overwhelmed by bad recommendations, user trust could also be fading over time. However, the dynamics of user trust depend heavily on user attitudes. For neutral users, good and bad recommendations weighed equally, which means a bad suggestion can be subtracted by a good one; for tolerant users, positive feedback has a greater impact than the bad one; for the intolerant users, a bad recommendation can affect user trust in a deeper level than the good one does (as shown in the figure below).

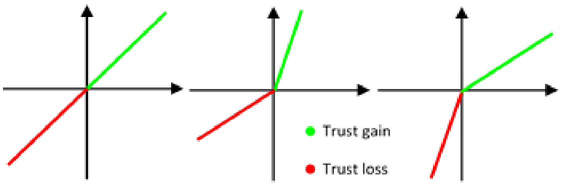

Figure 3: the relationships between user attitude and trust gain/ trust loss. (a) neutral users; (b) tolerant users; (c) intolerant users

To most users, a bad recommendation from the recommender system accounts for a more powerful impact on user trust than a good one because the users usually put high expectations on the recommender system, which is the reason why we need humans in the loop to leverage the accuracy of prediction/recommendation.

Improving the quality of the recommendations can be done by introducing human feedback into the system. Since most recommender systems collect usage or interest data, such as views, clicks or ratings from online users to make future predictions, it means that the system would narrow down the available data based on user feedback (Khenissi and Nasraoui, 2019). However, this could lead to the Filter Bubble (a state in which the users only see similar recommendations/information based on their preferences remembered by the algorithms, which might be over repetitive sometimes) problem in the long run.

Therefore, getting humans in an AI loop is not only to mitigate biases in the dataset but also to monitor and deal with the algorithms when facing the black box (the decision-making of AI that cannot be perceived or understood by humans). And with a trusting environment, where humans and AI built mutual trust (Huang et al., 2019), AI assistance can speed up the decision-making process. However, the core factor that determines the success or failure in decision-making by Human-Machine Team is the ability of humans to decide in which areas, and when to follow or to override the recommendations from AI. This implied that the algorithm only played a supporting role in the process to make it easier for humans to make the final selection only after reviewing the AI's inferences.

3.2 The Improvement of Task Performance

Artificial Intelligence (AI) has already been developed over time for the implementation in many cognitive application fields, such as language translation, visual recognition, voice recognition, and many others. AI potential of handling complex tasks in the unstructured environment derived from combining the cognitive skills of a human operator with autonomous tools and behaviors.

In this section, we mainly focus on 2 areas of applications:

1) pattern recognition, including visual recognition, speech recognition, language understanding, and other data labeling tasks, heavily relies on crowdsourcing approach;

2) cyber-physical systems, such as smart home, smart devices, and many other IoT innovation;

When tasks rely on human knowledge about context, it may be difficult to capture the entire necessary context in a way that a computer can reason about it (Cranor, 2008). Taking language translation for example, even though the AI has achieved a significant improvement in translating, there are still some gender biases when translating from one to another language (Srivastava and Rossi, 2019), where there is no differentiation for the third-person pronoun, and some misinterpretation caused by a misunderstanding from the context, which is more likely to happen with languages that are not widely used throughout the world. Thus, “borrowing” human intelligence in language understanding can indirectly improve the quality of translation.

For instance, Facebook posts created in foreign languages are usually automatically translated to the default language set by the user. However, the translation might not always be completely correct and up-to-date, especially on local slangs and dialects. Recently, Facebook decided to use Convolutional Neural Networks (CNNs) instead of the widely used Recurrent Neural Networks (RNNs) in order to potentially produce translations that more closely resemble localized texts. CNNs are faster in translating the whole sentence, and that is what Facebook is looking for since it is the giant communication platform, real-time delivering messages across the world is essential, yet real-time translating is more crucial for the non-English speaking users (Pascale, n.d.). Because Facebook is a free communication platform, what many users might subconsciously not realize is that we have constantly paid the price. By creating posts, comments, and other activities, we have already been generating a valuable data pool for Facebook's algorithms to learn, including the features where users can rate the translation by Facebook while being able to provide the suggested translation at the same time, which means we are part of the free knowledge transfer or free crowd workers. Thus, Facebook can effectively learn from online user behaviors and the changing language trends all around the world, then develop a better translation algorithm or technique.

Likewise, we might be familiar with Google reCAPTCHA (CAPTCHA is the abbreviation for Completely Automated Public Turing test to tell Computers and

Humans Apart), the Turing Test verification process we all go through to prove we are not robots online, occasionally popping up the “I’m not a robot” checkbox when having access to some websites. To ensure the website security and reduce user friction in the long run, Google reCAPTCHA generates a form of test that can only be solved by humans to run risk analysis. The test is generated in a form of texts, real words from the archived text that Optical Character Recognition (OCR) software has been unable to identify, where the first word given is recognizable while the second word is not. With the human contextualizing ability, only looking around the context will give humans the ideas of what the word should be. Hence, it is thought to be a win-win situation where Google can benefit from building its Google Books library while users can also be safe surfing on the websites. In a further study, (Hillen and Höfle, 2015) have extended to the Geo-reCAPTCHA approach, assessing the time and quality of the resulting geographic information. The result indicated that the users could solve the problem on average 19.2 seconds with an overall average accuracy of 82.2%, which has the potential to become a new data-rich geographic channel for crowdsourcing.

Worker engagement is not always depending on the monetized incentive. In the study of (Lee and Arora, 2019), they found that asking the questions people want to answer could lead to intensive data gathering at a much lower cost, and the dataset can be expanding by simply asking more questions. However, due to the noise caused by humans and the limited scale of samples, the quality of the results might not be as expected.

From the previous examples, these approaches can lead effectively to a larger data pool, and with humans’ efforts, it could be achieved in a shorter amount of time at a low cost or no cost. However, the low-paid or non-paid crowd workers’ efforts might not be consistent and always correct. Thus, there is still a need for hiring paid-crowd workers or human experts.

According to the case of the CrowdSynth effort on Galaxy Zoo (a crowdsourced astronomy project), the result illustrates that significant gains can be achieved from the optimization of access to human intelligence, which allows the AI to connect to crowd workers’ assistance (humans as AI helpers). At the same time, human input can also benefit agents in the process of learning how to act, where the AI can learn from human's feedback or suggestion (humans as AI teachers). The study shows that solving tasks with humans’ help can significantly improve the performance of traditional AI. By hiring only less than 50% of human workers during the process, the algorithm (CrowdSynth) has already reached its maximum accuracy in a consensus task (Kamar, 2016).

Based on the evaluation of the same case, with the condition of acquiring similar learning gains, it indicated that the approach of interactive teaching strategies lessened the amount of attention required from the teacher, compared to the teacher-initiated strategies.

By analogy, these strategies can apply to the other projects that are also hosted by the Zooniverse (a citizen science web portal). With the combination of active learning[1] and targeting specific volunteer groups allow machines and humans to each focus on their strengths, reducing the work for humans while improving the quality of training data for the models (Fortson et al., 2018). Nevertheless, optimizing the interaction of Human-in-the-loop is crucial.

To optimize the potential of Human-in-the-loop AI, it requires a comprehensive understanding of humans as helpers of AI systems as well as the way to access to human intelligence, including when to seek human help and when the systems can have access to human advice on how to act (Kamar, 2016). Optimizing Human-Machine Interaction (HMI) can advance the potential of the traditional AI, but there are some challenges that might be difficult to solve, which will be discussed further in section 4.

In the age of IoT, it is now feasible to ask your smart fridge about its contents and various properties of the food, including existence, count, category, and freshness (Gudovskiy et al., 2019). Cyber-Physical systems are broadly used around people in different age groups. Although the implementation of Human-in-the-loop AI in Cyber-Physical domains (the second category as described in section 2, where the system passively monitors the human and takes appropriate actions, including doing nothing), such as energy-saving infrastructure, smart home appliances, sleep tracking, health tracking, and others, has some limitations, we cannot deny the convenience and the real-time effect it has brought to our lives.

Nowadays, energy saving has become one of the biggest issues to be tackled in households. The study of (Das et al., 2019) proposed a novel graphical lasso based approach to perform the segmentation analysis by studying the feature correlations in a real-world energy social game dataset.

With slightly different approach, The Smart project can recommend users or act on their behalf, based on their past power consumption behavior, to achieve the goal of minimizing electricity wastage. Its approach aims at proactively recognizing the inhabitants' activities to conserve energy by implementing a smart home model. The entire flow of data is collected from the user behaviors and preferences consisting of the action and the actuation. After that, the device's power consumption values are clustered to determine device modes. The user actions have been modeled via transition matrix, whereas the need for optimal power consumption is taken care of by Reward Function[2]. Transition Probability Matrix and Reward Matrix are the key inputs and are to be computed in the Domain States, where HD denotes high-demand states and LD denotes low-demand states. The users' historical data are collected to predict the next state—if the current domain state and the prior differ, then the next action should be MOVE, otherwise, STAY. In this study, the MDP algorithm used the policy iteration algorithm where the agent chooses the best state using policy as a

[1] Active Learning refers to the process of learning, which allows data flow decisions to be made in near real-time (Fortson et al., 2018).

[2] Reward Function here refers to penalizing the domain states with higher power consumption and reward the domain states with lower power consumption (Verma et al., 2019).

rational agent should choose the action that maximizes its maximum utility. The domain states, LD and HD, and the MDP algorithm help maintain an effective trade-off with user-preferred states and states which consume less power. The result has proven that with the Smart Project, the energy could be saved up to 30% (Verma et al., 2019).

For elderlies and functionally locked-in individuals, Human-in-the-loop Cyber-Physical System (HiLCPSs) can be a game-changer to their lives. For example, SuperLimb (Supernumerary Robotic Limb), the pneumatically-driven robotic cane, is invented to actively assist elderlies during the sit-and-stand transition at the bedside, in the chair, and on the toilet. An inflatable vest with a depth sensor for ambient intention detection attached to the elderly’s body is made for human-robot interaction compliance. The result has proven that SuperLimb can effectively reduce lower limb efforts and elderly fall risks (Wu et al., 2020). For functionally locked-in individuals, who cannot interact with the physical world through their own movement and speech, an HiLCPS approach that merged that augments the neurophysiological capabilities of a functionally locked-in individual allows them to access the abilities of self-feeding, communication, mobility, and digital access (Schirner et al., 2013). Similarly, (Jain et al., 2020) applied supervised machine learning algorithms to model user engagement in the context of long-term, in-home SAR (Socially Assistive Robot) interventions for children with ASD. Therefore, it is proven that Cyber-Physical Systems can significantly improve their quality of lives by providing real-time assistance and reducing the needs of the healthcare system.

3.3 Humans and Complexity Reduction in Machine Learning

The essential benefit of Human-in-the-loop in Machine Learning (ML) is human feedback. Because the machines sometimes make mistakes, the feedback from humans allows AI system to learn and adapt throughout its life cycle.

In Reinforcement Learning, agents usually learn through feedback given in the form of Reward Function, which could be both rewarding correct actions and penalizing mistakes. The goal of the RL agent is to find the policy that maximizes the expected value of reward it receives in the long run, so an agent usually aims for the actions that can lead to the maximization in the discounted sum of future rewards. (Millán et al., 2019) compared the performance of the feedback for the continuous actor-critic algorithm and test the experiments in the cart-pole balancing task. The result has shown that the modeled human feedback can potentially increase the accumulated reward in comparison to the autonomous learning method.

Recently, many researchers have attempted to put the focus on interactive Machine Learning (a design of Machine Learning that supports and benefits from human interaction through interface) instead of traditional Machine Learning (ML). With Human-in-the-loop, interactive Machine Learning (iML) can benefit by human complementary abilities to solve computationally hard problems, such as subspace clustering, protein folding, or k-anonymization of health data, where human expertise can help to reduce an exponential search space through heuristic selection of samples. And by integrating humans in a learning phase, this can help reduce complexity in NP-hard problems (Holzinger, 2016).

In a state-of-the-art approach in interactive Machine Learning (iML), (Michael et al., 2020) has introduced a human analyst-based for defense applications where trust, safety, and quality are the main goals. The study found that interactive Machine Learning (iML) has the potential to improve both machine performance and user experience with autonomy. By integrating models of human cognition as feedback for Interactive Machine Learning (iML), the feedback from the human can directly address the shortcoming of the current iML model. Cognitive models, combined with self-reported data from surveys and physiological data, can provide a starting point for iML systems to optimize their suggestions for the overall performance of a Human-Machine Team.

In Machine Learning, feature selection is crucial for time-saving as it aims at minimizing the model's loss function by focusing on the most pertinent variables from a human perspective. In the study, (Correia and Lecue, 2019) have proven that the feedback from humans on feature selection can help improve the performance of the model. Moreover, this approach makes the decision-making more understandable for human users because the learning algorithm mimics human annotation in selecting the most relevant features, so the model can better reflect causal relationships in the experts' minds.

For the health domain, the research has proposed the Interactive Machine Learning (iML) approach to solve NP-hard problems. With Human-in-the-loop, the model can benefit from human integration, which can assist in computationally hard problem-solving by reducing an exponential search space through heuristic selection of samples, which accordingly can lead to a decrease in problem complexity (Holzinger et al., 2016). Moreover, the humans in the loop not only involved in the heuristic selection process but also interacting with the algorithm during the learning process to transform the black-box to a glass-box, also the Ant Colony algorithm is applied to solve the problem (Holzinger et al., 2017).

3.4 The Benefits of Co-Learning

In the previous sub-sections, we have already discussed humans in different roles, such as teachers, human experts, AI helpers, volunteers, and non-experienced workers (in the Crowdsourcing cases and others from section 3.1-3.3), as well as in different loops, open-loop, and closed-loop. In this session, we will be discussing humans in the role of the reciprocated partner of AI.

In the field of Hybrid Intelligence, a vast amount of studies has portrayed the role of humans as AI helpers or teachers with the responsibilities of giving feedback and correcting AI mistakes. However, it would be better if humans and AI can develop by learning from each other over time to grow as a team.

The Human-AI Co-learning model is built upon the principles of mutual understanding, mutual benefits, and mutual growth (Huang et al., 2019).

To create a mutual understanding in the collaboration, AI learns to explain the logic behind its action (this can be achieved by Explainable Artificial Intelligence, see more in section 4) while humans as well learn how to communicate their standard to AI. This can bridge the gap in the mismatch of both entities (see more in section 4.1), also knowing each other's strengths and weaknesses is the key to developing a common comprehension that leads to a mutual understanding.

After that, the processes of Self-learn & Reflect and Advice & Feedback will shape them into a better team driving towards superior outcomes, which cannot be achieved by each of them alone, and that is their mutual goal/mutual benefit.

Seeing that they are learning as a team with continuous feedback and consistent adaption, it means that they can form a common trust during the learning process, since then the mental models will be able to adapt to the changing environment, which leads to a mutual growth in both humans and AI.

Therefore, the traditional AI would benefit from getting humans into the AI loop as they can learn from each other and co-develop as a team to achieve superior results on the condition of a mutual understanding.

3.5 Conclusion

This section has completely answered the question of "How can Human-in-the-loop improve the performance of the traditional AI?", yet partially answered the question of "Does Human-in-the-loop AI perform better than each of them alone?".

Firstly, Human-in-the-loop AI can ensure that the system must be built in a way that conforms to AI ethics. Apart from that, Human-in-the-loop AI can enhance the security of life-related decision-making processes and primarily solve/reduce biases in the dataset injected into the system, which leads to the development of a fairer system. Moreover, integrating humans into the AI loop plays an important role in trust forming, and increase the credibility of the final results. Besides, humans can monitor the decision-making of AI and take appropriate actions when the system malfunctions.

Secondly, Human-in-the-loop AI can significantly improve the performance of cognitive tasks, such as data labeling, pattern recognition, and language translation, through the approach of crowdsourcing. Furthermore, humans and AI mutually benefit from Human-in-the-loop Cyber-Physical Systems. Humans can benefit from a real-time service that makes their lives easier while the AI benefits from real-time data feeding to the system, which can improve the performance of the AI over time.

Thirdly, humans can reduce the complexity in the Machine Learning process by providing feedback that allows the system to learn and adapt throughout its life cycle and by reducing an exponential search space through heuristic selection of samples.

Finally, co-learning activities allow humans and AI to develop as a team towards a mutual benefit—superior results that cannot be achieved by any of them alone, based on a mutual understanding. And this will lead to mutual growth in the long run as they can always learn from each other.

4. The Challenges of Human-in-the-loop

The purpose of Human-in-the-loop is to combine the strengths of both parties to overcome the limitation of each individual. However, this also implies that each of them has to adopt its counterpart's shortcoming, such as state of the human, human bias, AI rigidity, and other factors that can lead to human errors and machine failures. The biggest problem in Human-Machine Interaction is the mismatch between humans and AI mentality, and the adoption of the counterparts' weaknesses which has a correspondent impact on Human-AI interface. In order to create an effective Human-Machine Team, it is crucial to tackle human behavioral model problems.

4.1 The Mismatch

To overcome the traditional AI limitation, getting humans in AI Loop is the key. As described in section 3.2 and 3.4, humans are required to develop a comprehensive understanding of how the AI functions. At the same time, the AI should also reason about human partner's decision-making, which might need some intervention or guidance to direct its actions towards efficiency (Lee et al., 2019).

It is widely known that Human-AI Collaboration or the so-called Human-Machine Teaming (HMT) can enhance the performance of the traditional AI. However, the two dynamic, learning entities have distinct mental models (as described in 4.3), expertise, and abilities.

Thus, such fundamental differences or mismatches like the different aspects of the interaction, such as the powerful communicative impact of actions performed in a shared context, and the tradeoffs between performance gains and compatibility with existing human mental models, can lead to the unexpected failure and result in serious consequences.

Generally, updates to the software are to increase AI's predictive performance. However, it also simultaneously decreases the compatibility between the users and AI since the updates can lead to changes that are unfamiliar to the human users, which could harm team performance. In this case, the mismatch can also be perceived as incompatibility caused by new updates, which can be solved by penalizing new errors (Bansal et al., 2019a).

The mismatch between Humans and AI leads to the further problems:

1. When should AI ask for humans' helps?

2. When do AI have access to humans helps?

3. When/At which state in the workflow humans should add actions?

4. How to optimize the communication between humans and AI? (XAI)

5. How to smooth humans and AI Interaction? (Compatibility, penalizing new errors, tradeoff between performance and compatibility)

The biggest problems in Human-Machine Teaming, when to ask for help and when the human access is available, remain unclear and unanswered for the AI. There is no specific measurement on when exactly the AI should ask for human help. According to (Kamar, 2016), interactive teaching strategies (jointly initiated approach from both the student and the teacher, see more in (Amir et al., 2016)) can provide the same level of learning gains while requiring less attention from the human than the teacher-initiated strategies. Human assistance comes with the high cost and the performance of human workers strongly relies on the state of the human, so it means that the human should not be monitoring through the whole process and only interactively provide assistance when necessary, even though it is mostly impossible to specify the exact moment.

To address the question of "Where or when humans should add actions?", it is essential that we use human effort as efficiently as possible, and one significant loss is that humans waste effort adding actions at places/states that are not very important (Mandel et al., 2017). Thus, selecting a suitable state where adding actions can optimize the performance of the process or a state of which adding actions have a greater value than not adding action is the key. However, it is extremely difficult to determine where/at which state the next action should be added since humans are expected to possess expertise in various areas to understand the big scale of data.

Due to the characteristic mismatch of humans and AI, there might exist a communication problem in the workflow since they do not share a common language. To achieve the best quality of teamwork, humans and AI must be able to communicate with each other effectively. As human factors[3] fundamentally affect the efficiency of the process, it is essential to develop a clear perception of who the human is in the process to smooth the team effort.

At the same time, the black box of AI decision-making could also blur the logic behind the actions. In order to best exploiting Human-in-the-loop, AI needs to infer the intentions of the user implicitly through the observation of their actions and clearly communicate its own intentions through its own actions.

Therefore, AI should be able to explain their actions to humans so that humans can gain some insights into the mechanism of AI decision-making. Explainable AI (XAI) aims to improve various aspects of human-AI interaction, such as trust, traceability or predictability through explanations (Schrills and Franke, 2020).

Explainable AI (XAI) refers to Artificial Intelligence and Machine Learning techniques that can provide human-understandable justifications for their output behavior (Ehsan and Riedl, 2020). With the Explainable AI (XAI) approach, this can enhance and improve human-machine cognition (Preece et al., 2019). In a further study of Human-Centered Explainable AI (HCXAI), it is crucial to develop a clear perception of who the human in the loop is to further explain how the data is collected, what

[3] Human factors referred to the behavior and performance itself, including causes and effects of that behavior (Carroll and Olson, 1988).

data can be collected, and the most effective way of describing the why behind an action. Due to the sociotechnical terms, this approach, however, has to be sensitive to the values and the norms of the community as well as requiring active translational work from a diverse set of researchers to enhance the efficiency of the Human-Centered Explainable AI (HCXAI).

To develop a compatible or well-functioning Human-AI team, the machine needs a comprehensive understanding of the capabilities of its human partner, the cost and constraints associated when asking for help. At the same time, a successful partnership requires effort from the human as well. To understand how AI is functioning, the human needs to develop insights into the performance of the AI system, including its failures. Compatibility between humans and AI should be considered as a decisive variable in determining the performance of the teamwork.

The main problem of the compatibility of Human-Machine Teaming is that there exists a tradeoff between compatibility and performance. Once the system is updated, there will be a disruption in team performance. When the updates are applied to the system, the capability of AI prediction will be increasing while the compatibility of the team will suffer from unfamiliar changes, and it might take some time for both parties to adjust. In order to improve the compatibility after the update, the study proposed the re-training approach that penalize when the system makes new errors (Bansal et al., 2019b). This allows the update to maintain the accuracy while at the same time increase compatibility.

4.2 Human Error and Machine Failure

Solving complex problems by Human-in-the-loop has its pros and cons. In most cases, the collaborative approach outperforms each of the two entities separately. However, in some cases, the Human-Machine Teaming (HMT) produces results worse than either the human or machine would produce alone. This is because of the mismatch between the two entities as described in section 4.1 and the interfacing problem as described in section 4.3.

In particular, machine learning systems in the wild are facing difficulties with being adaptive to dynamic environments and self-adjusting, lack of what humans call common sense (Dellermann et al., 2019b) and gut feeling (Dellermann and Calma, 2018). In real-world problems, such as robotic execution, object manipulation or any other applications that have to deal with motion or physical movement, AI yet still does not perform as effectively as it is expected. Thus, it is common for AI to make errors or even cause failures because of its lack of commonsense, which will reflect in the system breakdown, financial loss or even loss of lives. There is a hidden gap lying in-between AI comprehension and real human intentions, so it is not easy for AI to interpret and progress the human request or feedback correctly in a timely manner.

Based on reasoning capabilities, it is difficult for the robot to self-reflect or realize when it does not function properly. Even when it realized that something went wrong, still it does not mean that it will be able to self-regulate or take action as soon as it happens. Besides, correcting robot failures is a time-consuming and costly process. Consequently, the H2R-AT (Human-to-Robot Attention Transfer) approach aims at transferring human intelligence to detect the anomalies of robot functioning in an early stage to solve the problems correspondingly to prevent further failures or malfunctions by monitoring the robot execution and transmitting verbal alerts when the abnormal actions occur (Song et al., 2020).

Due to the consideration of human factors, some studies show that it would be better to limit the human intervention in some tasks in the process or even completely leave out the humans because the state of the human can prone to failures or cause errors.

For example, in the healthcare application, the data-driven computational approaches started to gain popularity among clinical practice, such as medical image extraction, medical diagnosis, and others. CAI4CAI (Contextual Artificial Intelligence in Computer Assisted Interventions) approach aims at developing Human-AI Team for surgery tasks (Vercauteren et al., 2019), which requires a finer level of understanding of the surgical activities, and understanding of the ultimate Language of Surgery. Moreover, surgical gestures, surgical action, and surgical tool manipulation should be capture correctly to optimize team performance. However, humans are prone to fatigue and sometimes can cause miscommunications while machines are prone to failures. Thus, in this type of critical tasks requires a high level of specialization.

Similarly, in the security framework, humans often fail in their security roles. Whenever possible, secure system designers should find ways of keeping humans out of the loop to avoid human actions that are prone to failures (Cranor, 2008). The main factors that are more likely to cause failures consist of personal variables, intentions, and behaviors. Therefore, when designing a secure system that relies on humans, it is important to consider who these humans are likely to be and what their characteristics suggest about how they are likely to behave. It is also important to consider what relevant knowledge and experience these humans are likely to have to avoid both human and machine failures.

4.3 Human-AI Interface

According to the previous sections, humans and AI are two different entities that are merged as a team (Human-Machine Teaming or HMT). With their distinct mental models, there are some difficulties in user interfacing.

Mental models are crucially important to Human-Computer Interaction in the aspects of learning, memory, problem-solving, or planning. Normally, humans learn about AI systems through their lifetime experience, systematic training, and consistent imitation.

In this case, mental models refer to human’s behavior, the input-output characteristics of any software process run on a computer, or any information process mediated by people or machines (Carroll and Olson, 1988).

Being in the same team with a human, a robot/agent has to adjust its mental model or perception of the human teammate to achieve superior results as a team. Even with a sufficiently robust human-robot interface, robots will still need to understand and efficiently adapt to human behavior, much like humans adapt to the behavior of other humans. The design of Human-AI interface is a determination for the team performance, improper design of the autonomy could lead to the increase of human's cognitive load, the loss of situation awareness as a team, misaligned coordination, poorly calibrated trust, ultimately slow decision-making, deteriorated teaming performance, and even safety risks to the humans (Chakraborti et al., 2017). Therefore, the main challenge is how to develop representations of approximate and incomplete models that are easy to learn for human mental modeling and can support planning/decision-making for anticipating human behaviors. If the interface can reflect an appropriate model, this will help the user learn and develop a comprehensive understanding with less guidance and fewer errors made (Carroll and Olson, 1988).

The study of (Trautman, 2017) has shown that the failure in Human-Machine Teaming (HMT) is resulted by the deficiencies at the decision fusion level, which means by improving an individual entity might not reflect in the gains of team performance or reduce the chance of failures. Two entities should be able to develop together as a merged entity towards a common goal, and the fusion of two decision-makers should be as good as either in isolation.

Most of the existing Human-Machine Teaming (HMT) approaches do not have a robust mechanism to fuse human and machine information, which hinders the success of the Human-Machine Team that is able to perform greater than the sum of its parts.

4.4 Human Behavioral Model

According to the previous section, optimizing the communication between humans and AI requires a comprehensive perception of who the human is in the loop to understand how the framework should be modeled. From the AI perspective, the humans in the loop could be seen as helpers, experts, non-technical workers or any others. To further extend to a higher level, it is necessary to identify the complete spectrum of human-in-the-loop controls and identify the type of controls based on the role of the human in the loop as described in section 2.

From the cyber-physical system aspect, developing models of human behaviors is one of the biggest challenges so far. To be able to do so, it is important to answer the following questions:

1. which behaviors should be monitored?

2. how to properly model human behaviors by using the existing techniques?

3. How to effectively deal with human constantly changing behaviors?

These are open questions that need further research and not very easy to answer due to the complexity of human behaviors that cannot be parameterized.

Therefore, it is almost impossible to have an accurate measure for the human behavioral model since human behaviors can evolve over time according to the environment, sociological terms, physiological terms, and other variables.

Taking autonomous or semi-autonomous vehicles as an example, driver modeling is crucial as human factors are inevitable in the workflow. However, in such high-stake decision-making domains like autonomous driving, it is important to consider the individuality or characteristics of each individual. (Albaba and Yildiz, 2020) stated that different levels of reasoning exist for different humans, so acquiring appropriate behavior model for the individuals is crucial.

Due to human factors, the autonomous agent has to adapt itself to function properly in a shared control framework. The study of (Luo et al., 2020) indicated that the adaptive haptic control scheme resulted in a significantly lower workload, higher trust in autonomy, better driving task performance and smaller control effort.

Apart from the driver behavior models, another factor to be considered is pedestrian models. Pedestrians are active agents with complex interactive motions, so predicting individual pedestrians' likely destinations and trajectories is necessary yet difficult to achieve. One of the biggest challenges is that model development requires an interdisciplinary approach, including simple visual models for the detection of pedestrians and predicting future movements using psychological and sociological methods. Besides, it requires autonomous vehicles to utilize many very different levels of pedestrian models, each addressing different aspects of perception and action (Camara et al., 2020).

4.5 Conclusion

In this section, we have discussed The Mismatch between humans and AI, which is one of the biggest obstacles for Hybrid Intelligence performance optimization. Combining two different entities means that the pros and cons of each are as well integrated, which reflect in Human Errors and Machine Failures that can also be caused by the imperfect Human-AI Interface and solved by Human Behavioral Model, but modeling human behaviors is still far from being perfect.

Why does getting humans into the AI loop become challenging and costly? First of all, the characteristic mismatch between humans and AI has the biggest impact on this challenge. Because such fundamental differences or mismatches like the different aspects of the interaction, such as the powerful communicative impact of actions performed in a shared context, and the tradeoffs between performance gains and compatibility with existing human mental models, can lead to the unexpected failure and result in serious consequences. And it would become costly since there are no single criteria to determine where to add actions, so most of the time people might be wasting time, effort, and budget in the wrong place.

Secondly, integrating humans into the AI loop means that both entities would absorb each other's strengths and weaknesses. Hence, human errors and machine failures are introduced to the framework.

Thirdly, the perfect Human-AI interface might not be possible to achieve since the characteristic mismatch has rooted for the biggest challenge. Besides, the improper interface could result in worse outcomes than each of them in isolation.

Finally, the human behavioral model has its limitations. Since it is not easy to indicate which behavior to model and human behaviors are constantly changing over time, it is very difficult to model human behaviors appropriately.

5. The Future of Human-in-the-loop

Since AI has been developing consistently over the past decades and has overcome some of its limitations in several aspects and significantly surpassed human abilities in many areas and started to take over human roles in financial, medical, juridical and others, the AI applications have rooted in our daily lives over time. However, AI is still far from being perfect, but with human assistance or so-called human-in-the-loop AI, its potential abilities will become far more competitive than the traditional AI.

The direction of the future AI will be developing towards Artificial General Intelligence (AGI) instead of the narrow AI that is commonly implemented in these few decades. Therefore, AI developers have to overcome the existing limitations of the current AI by integrating humans into the workflow, which can lead to the enhancement of Human-AI collaboration, but somehow a new challenge.

The key challenges of developing human-in-the-loop AI are the following:

1. How to put human-in-the-loop in practice to the real world problems?

2. How to improve the quality in user interface?

3. How to reduce the mismatch between humans and AI?

Developing well-functioning human-in-the-loop AI requires a paradigm shift from hybrid systems to hybrid teamwork, where humans and AI are equally-positioned partners. Henceforth, it requires deeper reasoning capabilities for machines to make decisions not only about how they are accomplishing their tasks but also about how they can support their teammates towards the success of the collaborative activity (Kamar, 2016).

Applications that are able to deal with real-world problems require a continuously collaborating socio-technological ensemble integrating humans and machines (Dellermann et al., 2019a).

Due to AI's limitations in reasoning, object manipulating, natural language processing (NLP), and others, there is a certain difficulty in building a perfect Human-Machine Team (HMT), especially on the interface level in both physical tasks and cognitive tasks. To reduce the mismatch between humans and AI, developing compatible mental models, and researching on Explainable AI (XAI) is crucially important because it is necessary to form a trusting environment, where the team members can fully trust each other.

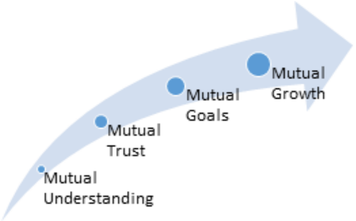

Figure 4: Human-in-the-loop development based partially on (Huang et al., 2019)

In developing human-in-the-loop AI, trust is the main factor deciding whether team performance can achieve an optimal state or not.

First, humans and AI need to build a common ground, which is a comprehensive understanding of each other mental models and actuations. With black-box approaches, the algorithm usually cannot explain why and how the decision has been made, which leads to the lack of transparency, and this can cause a tremendous impact on trust and acceptance among end-users. Thus, AI has to be able to explain the logic behind its actions, behaviors or decisions to humans so that humans can gain some insights about how AI perceives and processes input data. At the same time, AI should be able to adapt to human behaviors and mental models that differed based on who the human is in the loop (different backgrounds, expertise, ages, and other characteristics). It is also important to agree on a common perception of human roles so that both parties can fine-tune to enhance compatibility and effectively collaborate.

Based on the mutual understanding, the trust has gradually been forming within the team, which drives the team towards the same direction. Having a common goal can help both humans and AI to understand the clear positions and roles of each team member as well as to motivate them to achieve a shared mission.

To create reciprocity between both parties, humans should not be involved solely in the pre-training process or teaching process. Instead, humans and AI should learn and grow together as coalitions to leverage the pros and cons of each equally.

Most importantly, AI should only play a supporting role in the process to make it easier for humans to make the final decision. Thus, humans should be able to decide in which areas, and when to follow or to override the recommendations from AI.

6. Summary

The core of Hybrid Intelligence is the combination of human and machine intelligence. Hybrid Intelligence takes human expertise and intention into account when making meaningful decisions and taking appropriate actions conforming to ethical, legal, and societal values. The key design of Hybrid Intelligence is getting humans into the AI loop, which aims to enhance the performance of the traditional AI.

Although there are various types of tasks that AI performs better than humans, still some tasks cannot be achieved by AI alone, which is why we need Human-in-the-loop AI to overcome these limitations.

Ethics, security, fairness, and trust are the most decisive factors in determining whether the results produced by Human-Machine Team (HMT) are reliable or not. Human-in-the-loop AI can enhance the security of high-stake decision-making processes and primarily solve/reduce biases in the dataset injected into the system, which leads to a fairer system. Moreover, integrating humans into the AI loop plays an important role in trust forming, and increase the credibility of the final results. Besides, humans can monitor the decision-making of AI and take appropriate actions when the system malfunctions. Moreover, Human-in-the-loop AI can significantly improve the performance of cognitive tasks through the approach of crowdsourcing. Furthermore, humans and AI mutually benefit from Human-in-the-loop Cyber-Physical Systems, which means that humans can benefit from a real-time service while the AI benefits from real-time data feeding to the system. Apart from that, humans can reduce the complexity in the Machine Learning process by providing feedback that allows the system to learn and adapt throughout its life cycle and by reducing an exponential search space through heuristic selection of samples. Besides, co-learning activities allow humans and AI to develop as a team towards a mutual benefit—superior results that cannot be achieved by any of them alone, based on a mutual understanding. And this will lead to mutual growth in the long run as they can always learn from each other.

However, the mismatch between humans and AI is one of the biggest obstacles for Hybrid Intelligence performance optimization. Because such fundamental differences or mismatches like the different aspects of the interaction, such as the powerful communicative impact of actions performed in a shared context, and the tradeoffs between performance gains and compatibility with existing human mental models, can lead to the unexpected failure and result in serious consequences. Combining two different entities means that the pros and cons of each are as well integrated, which reflect in Human Errors and Machine Failures that can also be caused by the imperfect Human-AI Interface. The perfect Human-AI interface might not be possible to achieve since the characteristic mismatch has rooted for the biggest challenge. Besides, the improper interface could result in worse outcomes than each of them in isolation. But this problem and can be solved by Human Behavioral Model. Unfortunately, since it is not easy to indicate which behavior to model and human behaviors are constantly changing over time, it is very difficult to model human behaviors appropriately.

To reduce the mismatch between humans and AI, developing compatible mental models, and researching on Explainable AI (XAI) is crucially important because it is necessary to form a trusting environment, where the team members can fully trust each other.

First, humans and AI need to build a common ground, which is a comprehensive understanding of each other mental models and actuation. It is also important to agree on a common perception of human roles so that both parties can fine-tune to enhance compatibility and effectively collaborate.

Based on the mutual understanding, the trust has gradually been forming within the team, which drives the team towards the same direction. Having a common goal can help both humans and AI to understand the clear positions and roles of each team member as well as to motivate them to achieve a shared mission.

To create reciprocity between both parties, humans should not be involved solely in the pre-training process or teaching process. Instead, humans and AI should learn and grow together as coalitions to leverage the pros and cons of each equally.

Most importantly, AI should only play a supporting role in the process to make it easier for humans to make the final decision. Thus, humans should be able to decide in which areas, and when to follow or to override the recommendations from AI.

Reference

Books

Lee, K.-F., 2018. AI superpowers: China, Silicon Valley, and the new world order. Houghton Mifflin Harcourt, Boston.

Luger, G., Stubblefield, W., 1998. Artificial Intelligence (3rd ed., pp. 17-18). Addison Wesley Longman, Inc.

Ridley, D., 2012. The Literature Review--A Step-By-Step Guide for Students. 2nd ed.

Articles

Albaba, B.M., Yildiz, Y., 2020. Driver Modeling through Deep Reinforcement Learning and Behavioral Game Theory. arXiv:2003.11071 [cs].

Amir, O., Kamar, E., Kolobov, A., Grosz, B., 2016. Interactive Teaching Strategies for Agent Training. In Proceedings of IJCAI 2016.

Bansal, G., Nushi, B., Kamar, E., Weld, D., Lasecki, W., Horvitz, E., 2019. A Case for Backward Compatibility for Human-AI Teams. arXiv:1906.01148 [cs, stat].

Bansal, G., Nushi, B., Kamar, E., Weld, D.S., Lasecki, W.S., Horvitz, E., 2019. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. AAAI 33, 2429–2437.

Camara, F., Bellotto, N., Cosar, S., Weber, F., Nathanael, D., Althoff, M., Wu, J., Ruenz, J., Dietrich, A., Markkula, G., Schieben, A., Tango, F., Merat, N., Fox, C.W., 2020. Pedestrian Models for Autonomous Driving Part II: high level models of human behaviour. IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS 34.

Carroll, J. M., Olson, J. R., 1988. Mental models in human-computer interaction. In Handbook of human-computer interaction (pp. 45-65). North-Holland.

Chakraborti, T., Kambhampati, S., Scheutz, M., Zhang, Y., 2017. AI Challenges in Human-Robot Cognitive Teaming. arXiv:1707.04775 [cs].

Correia, A.H.C., Lecue, F., 2019. Human-in-the-Loop Feature Selection. AAAI 33, 2438–2445.

Cranor, L.F., 2008. A framework for reasoning about the human in the loop. In Proceedings of the 1st Conference on Usability, Psychology, and Security (UPSEC'08). USENIX Association, USA, Article 1, 1–15.

Das, H.P., Konstantakopoulos, I.C., Manasawala, A.B., Veeravalli, T., Liu, H., Spanos, C.J., 2019. A Novel Graphical Lasso based approach towards Segmentation Analysis in Energy Game-Theoretic Frameworks. 1702-1709. 10.1109/ICMLA.2019.00277.

Dellermann, D., Calma, A., 2018. Making AI Ready for the Wild: The Hybrid Intelligence Unicorn Hunter. SSRN Journal.

Dellermann, D., Calma, A., Lipusch, N., Weber, T., Weigel, S., Ebel, P., 2019. The Future of Human-AI Collaboration: A Taxonomy of Design Knowledge for Hybrid Intelligence Systems. In T. Bui (ed.), HICSS (p./pp. 1-10). ScholarSpace.

Dellermann, D., Lipusch, N., Ebel, P., Leimeister, J.M., 2019. Design principles for a hybrid intelligence decision support system for business model validation. Electron Markets. https://doi.org/10.1007/s12525-018-0309-2.

Driggs-Campbell, K., Shia, V., Bajcsy, R., 2015. Improved driver modeling for human-in-the-loop vehicular control. In 2015 IEEE International Conference on Robotics and Automation (ICRA) (pp. 1654-1661). IEEE.

Ehsan, U., Riedl, M.O., 2020. Human-centered Explainable AI: Towards a Reflective Sociotechnical Approach.

Elish, M. C., 2019. Moral crumple zones: Cautionary tales in human-robot interaction. Engaging Science, Technology, and Society, 5, 40-60.

Fortson, L., Wright, D., Lintott, C., Trouille, L., 2018. Optimizing the Human-Machine Partnership with Zooniverse.

Gudovskiy, D., Han, G., Yamaguchi, T., Tsukizawa, S., 2019. Smart Home Appliances: Chat with Your Fridge.

Hillen, F. and Höfle, B., 2015. Geo-reCAPTCHA: Crowdsourcing large amounts of geographic information from earth observation data. International Journal of Applied Earth Observation and Geoinformation. 40.29-38.10.1016/j.jag.2015.03.012.

Holzinger, A., 2016. Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics 3, 119–131.

Holzinger, A., Plass, M., Holzinger, K., Crisan, G.C., Pintea, C.-M., Palade, V., 2017. A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop.

Holzinger, A., Plass, M., Holzinger, K., Crişan, G.C., Pintea, C.-M., Palade, V., 2016. Towards interactive Machine Learning (iML): Applying Ant Colony Algorithms to Solve the Traveling Salesman Problem with the Human-in-the-Loop Approach, in: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (Eds.), Availability, Reliability, and Security in Information Systems. Springer International Publishing, Cham, pp. 81–95.

Huang, Y.-C., Cheng, Y.-T., Chen, L.-L., Hsu, J.Y., 2019. Human-AI Co-Learning for Data-Driven AI.

Jain, S., Thiagarajan, B., Shi, Z., Clabaugh, C., Matarić, M.J., 2020. Modeling Engagement in Long-Term, In-Home Socially Assistive Robot Interventions for Children with Autism Spectrum Disorders. Sci. Robot. 5, eaaz3791.

Kamar, E., 2016. Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence. In IJCAI (pp. 4070-4073).

Khenissi, S., Nasraoui, O., 2019. Modeling and Counteracting Exposure Bias in Recommender Systems. Paper 3182.https://doi.org/10.18297/etd/3182.

Kleinberg, J., Mullainathan, S., Raghavan, M., 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In 8th Innovations in Theoretical Computer Science Conference (ITCS 2017). Schloss Dagstuhl-Leibniz-Zentrum fuer nformatik.

Klute, F., Li, G., Löffler, R., Nöllenburg, M., Schmidt, M., 2019. Exploring Semi-Automatic Map Labeling. In 27th ACM SIGPATIAL International Conference on Advances in Geographic Information Systems (SIGPATIAL '19), November 5-8, 2019, Chicago, IL, USA. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3347146.3359359.

Lee, G., Mavrogiannis, C., Srinivasa, S.S., 2019. Towards Effective Human-AI Teams: The Case of Collaborative Packing. arXiv:1909.06527 [cs].

Lee, J.J., Arora, S., 2019. A Free Lunch in Generating Datasets: Building a VQG and VQA System with Attention and Humans in the Loop.

Leslie, D., 2019. Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute. https://doi.org/10.5281/zenodo.3240529.

Luo, R., Weng, Y., Wang, Y., Jayakumar, P., Brudnak, M.J., Paul, V., Desaraju, V.R., Stein, J.L., Ersal, T., Yang, X.J., 2020. A Workload Adaptive Haptic Shared Control Scheme for Semi-Autonomous Driving.

Mandel, T., Liu, Y.-E., Brunskill, E., Popovic, Z., 2017. Where to Add Actions in Human-in-the-Loop Reinforcement Learning. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17). AAAI Press, 2322–2328.

Michael, C.J., Acklin, D., Scheuerman, J., 2020. On Interactive Machine Learning and the Potential of Cognitive Feedback. CoRR abs/2003.10365.

Millán, C., Fernandes, B., Cruz, F., 2019. Human feedback in continuous actor-critic reinforcement learning. Computational Intelligence. In ESANN 2019: Proceedings, 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, ESANN, [Bruges, Belgium], pp. 661-666.

Munir, S., Stankovic, J.A., Liang, C.-J.M., Lin, S., 2013. Cyber physical system challenges for human-in-the-loop control. In Presented as Part of the 8th International Workshop on Feedback Computing.

Pelta, D.A., Verdegay, J.L., Lamata, M.T., Corona, C.C., 2020. Trust dynamics and user attitudes on recommendation errors: preliminary results. arXiv:2002.04302 [cs].

Peng, A., Nushi, B., Kiciman, E., Inkpen, K., Suri, S., Kamar, E., 2019. What You See Is What You Get? The Impact of Representation Criteria on Human Bias in Hiring. In AAAI Conference on Human Computation and Crowdsourcing. 125-134.

Pessach, D., Shmueli, E., 2020. Algorithmic Fairness. arXiv:2001.09784 [cs, stat].

Preece, A., Braines, D., Cerutti, F., Pham, T., 2019. Explainable AI for Intelligence Augmentation in Multi-Domain Operations. arXiv:1910.07563 [cs].

Schirner, G., Erdogmus, D., Chowdhury, K., Padir, T., 2013. The future of human-in-the-loop cyber-physical systems. In Computer, vol. 46, no. 1, pp. 36-45.

Schrills, T., Franke, T., 2020. How to Answer Why -- Evaluating the Explanations of AI Through Mental Model Analysis. arXiv:2002.02526 [cs.HC].

Smith, C.J., 2020. Designing Trustworthy AI: A Human-Machine Teaming Framework to Guide Development. Carnegie Mellon University. Conference contribution. https://doi.org/10.1184/R1/12119847.v1.

Song, B., Peng, Y., Luo, R., Liu, R., 2020. Human-to-Robot Attention Transfer for Robot Execution Failure Avoidance Using Stacked Neural Networks. arXiv:2002.04242 [cs].

Srivastava, B., Rossi, F., 2019. Towards Composable Bias Rating of AI Services. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18). Association for Computing Machinery, New York, NY, USA, 284–289. DOI:https://doi.org/10.1145/3278721.3278744.

Trautman, P., 2017. A Mathematical Theory of Human Machine Teaming.

Vercauteren, T., Unberath, M., Padoy, N., Navab, N., 2020. CAI4CAI: The Rise of Contextual Artificial Intelligence in Computer Assisted Interventions. Proc. IEEE 108, 198–214.

Verma, M., Bhambri, S., Buduru, A.B., 2019. Making Smart Homes Smarter: Optimizing Energy Consumption with Human in the Loop. arXiv:1912.03298 [cs].

Webster, J., Watson, R.T., 2002. Analyzing The Past to Prepare for The Future: Writing a Literature Review.

Wu, X., Liu, H., Liu, Z., Chen, M., Wan, F., Fu, C., Asada, H., Wang, Z., Song, C., 2020. Robotic Cane as a Soft SuperLimb for Elderly Sit-to-Stand Assistance.

Zanzotto, F.M., 2019. Viewpoint: human-in-the-loop artificial intelligence. J. Artif. Int. Res. 64, 1 (January 2019), 243–252. DOI:https://doi.org/10.1613/jair.1.11345.

Website

Baraniuk, C., 2015. The Cyborg Chess Players That Can'T Be Beaten. [online] BBC. Available at: <https://www.bbc.com/future/article/20151201-the-cyborg-chess-players-that-cant-be-beaten> [Accessed 1 December 2019].

Cindicator, 2017. Hybrid Intelligence For Effective Asset Management. White Paper Version 1.2.16. [online] Available at: <https://cdn.cindicator.com/c338a33c-b654-4273-8dc8-a147c1d05f2d/-/inline/yes/Cindicator_WhitePaper_en.pdf> [Accessed 1 December 2019].

Deahl, D., 2017. China Launches Cyber-Court To Handle Internet-Related Disputes. [online] The Verge. Available at: <https://www.theverge.com/tech/2017/8/18/16167836/china-cyber-court-hangzhou-internet-disputes> [Accessed 1 December 2019].

Pascale, R., n.d. Google Vs. Facebook Machine Translation – What Is The Difference? – Venga Global. [online] Venga Global. Available at: <https://www.vengaglobal.com/blog/google-vs-facebook-nmt-difference> [Accessed 30 March 2020].