The Effect of Level of AI Transparency on Human-AI Teaming Performance Including Trust in Machine Learning Interface

by

GeeBeum Park

**A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science
(Human Centered Design and Engineering)
in the University of Michigan – Dearborn
2023**

Master's Thesis Committee:

**Associate Professor Sang-Hwan Kim, Chair
Assistant Professor Areen Alsaid
Assistant Professor Junho Hong**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF APPENDICES

ABSTRACT

The objective of this study is to investigate the impact of various levels of transparency in a human-AI teaming task on human performance, including task performance, trust, and workload.

A task simulator using real-time AI was developed and used to compare two different levels of information transparency in AI. A total of 20 participants participated in the experiment, and each participant was asked to play a Pictionary game by drawing given words while the AI presented a guess of the words, in a simple form of human-AI cooperation. The task performance was measured for two different levels of transparency for displaying the top 1 or top 5 objects that the AI recognized as being the most similar to the participant's drawing. During the experiment, task completion time, the number of errors, an eye movement profile, subjective workload, and subjective ratings on trust were collected and analyzed, along with this post-trial interview.

Results revealed that participants paid more attention to information display under conditions of higher-transparency condition while ameliorating workload and increasing the level of trust in cooperating with AI. Interview results identified the importance of individual differences in HAT performance, and as suggestions in providing transparency along with explainability information.

While the study includes limitations such as limited levels of transparency, it confirms the benefits of transparency and other human factors issues in HAT. It is expected that the study can serve as a basis for further studies to determine effective transparency in HAT.

CHAPTER 1: INTRODUCTION

1.1 AI

In recent decades, as interest in AI has increased, AI technology has advanced at a breakneck pace. Numerous start-up companies and IT giants are making significant investments to enhance AI technology, according to Lu et al. (2020). Amazon applied artificial intelligence in its delivery systems for autonomous robots. Facebook has also built ‘DeepFace,’ an artificial intelligence-based face recognition system. Microsoft has developed ‘ChatGPT’, AI is increasingly becoming applicable in a broader range of fields. Dwivedi et al. (2021) stated that AI technology are being actively used in the following areas: digital imaging, education, government, healthcare, manufacturing, robotics, and supply chain. There have been investigations performed in the industries of manufacturing, construction, and production on the use of intelligent agents to build automated systems or to monitor and manage processes. (Muhuri er al., 2019; Parveen, 2018; Yang et al., 2017; Zhong et al., 2017). AI is being applied to enhance patient care, disease diagnosis, and patient services in the healthcare industry. The healthcare industry is utilizing AI to improve patient care, disease diagnosis, and patient services (Khanna et al., 2013; Dreyer and Allen, 2018; Houssami et al., 2017). Several kinds of research have shown that the application of AI in education increases teacher effectiveness, student engagement, and library system efficiency (Chaydhri et al., 2013; Arlitsch and Newell, 2017).

The AI technology trends that have received the most attention recently are autonomous driving technology, speech recognition, recommendation systems, natural language processing, and image recognition. Continuous research has been undertaken in autonomous driving to build driving decision and control algorithms that can drive safely and smoothly in a variety of road situations by identifying the surrounding environment using radar and cameras (Bojarski et al. 2016; Kiran et al., 2021). In the field of speech recognition, various research on speech processing techniques and language modeling with deep learning models is ongoing. (Chan et al., 2016; Amodei et al., 2016). Furthermore, intensive research on sequential recommendation, collaborative filtering-based recommendation system, commonsense reasoning, image segmentation, and few-shot learning is driving AI technology forward.

1.2 Human-AI Teaming

However, due to the rapid advancement of AI technology, humans are confronting a society in which they should coexist with AI. AI software is capable of rapidly processing massive amounts of data for trained occurrences or circumstances. However, on the other side, AI will remain insufficient for identifying and operating in the most unexpected instances in the upcoming future (National Academies of Sciences, Engineering, and Medicine, 2021). Therefore, human intervention is necessary for addressing this challenge, giving rise to the concept of 'human-AI teaming,' in which humans and artificial intelligence support each other and collaborate.

A human-AI team is described as “one or more individuals and one or more AI systems that require collaboration and coordination to complete a task successfully” (Cuevas et al., 2007). This concept is founded on human-autonomy teams (HAT), which were described by several researchers in the 1990s (O’Neill et al., 2022). The human-autonomy team is a military model created at the beginning of the 2000s by the US Department of Defense to improve the interaction between people and unmanned systems. It was quickly adopted not just by the military, but also by a wide range of other industries, including aviation operations, driving, industrial automation, and healthcare. As artificial intelligence has developed, it has evolved into a human-AI team (Endsley, 2017). Human-AI teaming has the potential to improve decision-making and team performance in high-risk, time-critical environments by leveraging the complementary capabilities of humans and AI (Caldwell et al., 2022). Effective human-AI team can enhance situation awareness, decision support, and coordination between teams, leading to better decision-making, increased productivity, reduced errors, and the ability to solve complex problems.

However, the U.S. National Academy of Sciences, Engineering, and Medicine (2021) discovered that designing a successful human-AI teaming has some significant challenges in sub-fields such as situational awareness (SA), human-AI interaction, and trust. When humans and AI collaborate as a team, there is a need for team SA, which each team member requires for their individual roles, as well as shared SA, which all team members should have (Bolstad et al., 2002; Endsley and Jones, 2001). Therefore, it is critical to show information based to each individual's SA requirements in order to avoid overload (Bolstad and Endsley, 1999). Also, further study is required to improve human SA and shared SA for AI systems (USAF, 2015).

Human-AI team interaction occurs when humans and artificial intelligence systems collaborate and communicate with one another. This interaction can vary in different ways, such as AI making decisions and humans examining and modifying them. Since humans and AI have distinct abilities and limitations, it is essential to investigate how they can complement each other's strengths through collaboration (National Academies of Sciences, Engineering, and Medicine, 2021). Therefore, research in improving coordination between humans and AI for shared tasks, as well as maintaining situational awareness when working with highly automated

AI systems, is presently being undertaken (Endsley, 2017; Onnasch, Wickens, and Manzey, 2014; Wickens, 2009).

The success of human-AI teams relies heavily on trust, especially in multi-domain operations where quick and efficient information processing, filtering, and communication are necessary for effective decision-making (National Academies of Sciences, Engineering, and Medicine, 2021). To ensure this success, it is essential to understand how trust is influenced by organizational and social contexts (Gao, Lee, and Zhang, 2006). Achieving this understanding requires restructuring the research, testing, and evaluation of trust, including the definition of trust metrics and evaluation environments. While past research has focused on increasing human reliance on automation through trust calibration (Lee and See, 2004), it has not fully addressed related issues. As a result, further research on trust in human-AI teaming is necessary.

1.3 Transparency

AI systems that are sufficiently transparent in their functioning to allow for effective human interaction and supervision are required (Shively et al., 2017). Therefore, numerous research on the transparency of AI systems has been conducted. Endsley et al. (2003) defined AI system transparency as the system's understandability and predictability. There are two interconnected components of AI system transparency, which are display transparency and explainability. Display transparency, as part of situation awareness, gives a real-time understanding of the AI system's behaviors. (National Academies of Sciences, Engineering, and Medicine, 2021). Display transparency has been demonstrated to be beneficial in terms of enhancing team performance, situation awareness, and trust calibration. Explainability brings information on the logic or reasoning of the AI system's actions or suggestions in a backward-looking manner. AI explainability has been found to increase trust.

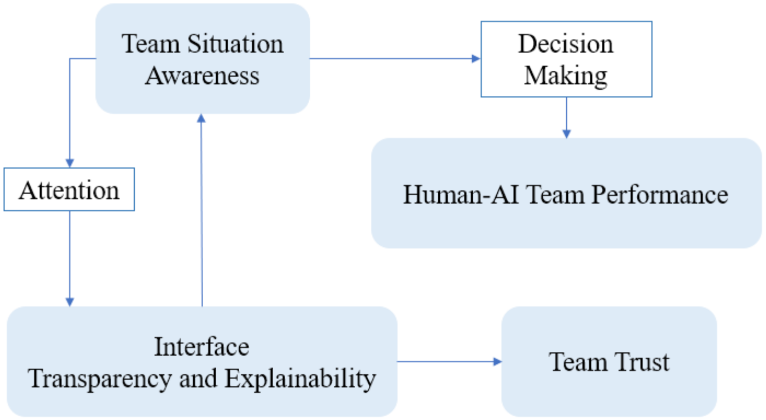

Team Situation Awareness
Decision Making
Attention
Human-AI Team Performance
Interface Transparency and Explainability
Team Trust

Figure 1. Relationship between transparency of AI, situation awareness, trust, and team performance.

Of these two elements, this study focuses on display transparency. Display transparency, as part of situation awareness, gives a real-time understanding of the AI system's behavior (National Academies of Sciences, Engineering, and Medicine, 2021). Display transparency aims to help the user maintain a clear understanding of the system and its environment without getting overwhelmed (Mercado et al., 2016). Previous studies have shown that display transparency improves the oversight of automation and performance, situational awareness, performance, and the capacity to develop trust (Bass, Baumgart, and Shepley, 2013; Bean, Rice, and Keller, 2011; Stowers et al., 2017; Boyce et al., 2015; Chen et al., 2014b; Schmitt et al., 2018; Selkowitz, Lakhmani, and Chen, 2017; Hoff and Bashir, 2015; Panganiban, Matthews, and Long, 2020;). In contrast, other studies found that increasing transparency can lead to a greater workload, especially when there is an excessive amount of information that can have a detrimental effect on performance (Kunze et al., 2014a; Chen et al., 2018).

In this regard, studies have been carried out to determine the effect of transparency on operator performance, trust, workload, and user confidence. Mercado et al. (2016) aimed to investigate the effects of agent transparency on operator performance, trust, and workload in the context of human-agent teaming for multirobot management. The researchers designed a simulated heterogeneous multi-UxV planning task, where operators had to decide on the appropriate action to be carried out by the UxV based on various factors such as the commander's intent, vehicle capability, and environmental constraints. The study discovered that improving agent transparency in a multi-UxV management task improved overall team performance and trust calibration. Additionally, increasing transparency did not result in added speed or workload costs, and operators found a transparent agent more trustworthy and usable.

Yang et al. (2017) studied how trust evolves over time in human-automation interaction and how transparency affects moment-to-moment changes in trust. The study aimed to provide insights into the dynamic nature of trust in military reconnaissance scenario. Participants took part in an experiment in controlling remote robots with automated threat detectors. They discovered that trust evolved and stabilized with experience, and that automation transparency influenced momentary changes in trust. Another research by Wright et al. (2019) studied the impact of transparency and reliability of a robotic agent on user confidence and perceived reliability in a simulated military environment. The study involved 84 participants interacting with a robot that had varying levels of transparency and reliability. They found that transparency did not affect participants' performance, workload, or situational awareness, while reliability had a significant impact on participant trust and perceptions of the robot. Although studies have shown that increasing transparency levels can lead to increased trust, there is still a need for further research to understand how transparency affects participants' trust, performance, situational awareness, and workload. Additionally, more research is necessary to determine the appropriate amount of information that should be given to participants.

1.4 Research Objective

The impact of transparency varied across different studies. Some studies found that increased transparency improved trust and performance, while having no effect on workload (Mercado et al., 2016; Selkowitz et al., 2017). On the other hand, other studies showed that while transparency improved trust and performance, it had negative effects on workload and performance time (Chen et al., 2018; Wright et al., 2016), resulting in a trade-off between the advantages and disadvantages of transparency. Moreover, there was a lack of research on how much information about transparency should be provided. Therefore, the objective of this research was to examine the effects of different levels of transparency on trust, performance, and workload. An attempt was also made to determine a level of transparency that does not result in trade-offs. Furthermore, this study discovered other critical human factors issues that accompany human-AI teaming tasks through an experiment.

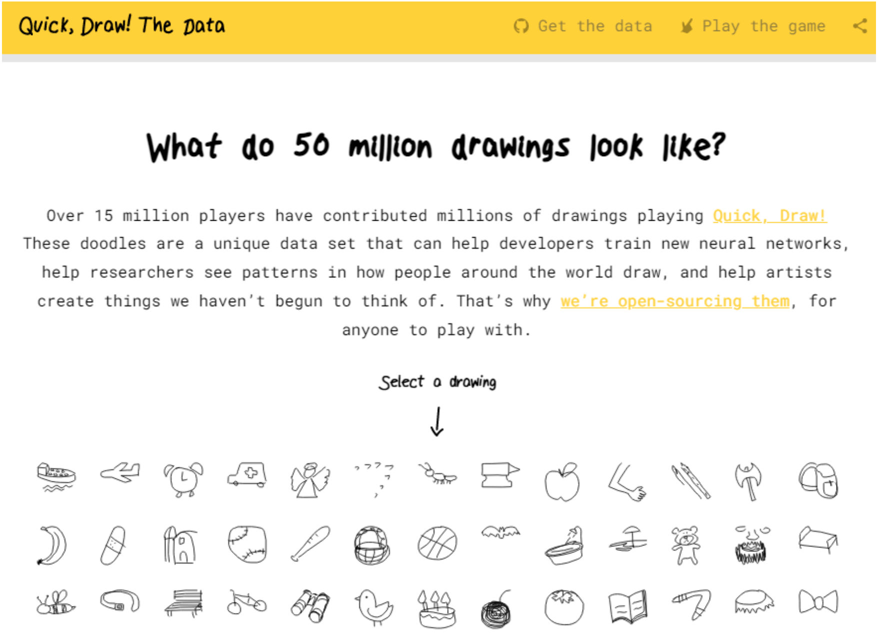
CHAPTER 2: METHODS

`As stated earlier, this study aimed to investigate the impact of various levels of transparency on trust, performance, and workload. To investigate this, an experiment was carried out that imitated a Pictionary game, where one person draws an image, and another person tries to guess what it is. As a participant sketched drawings during the experiment, an AI system attempted to predict what the images were. Then, the system provided transparency information on the side. Previous studies have shown that offering clues is more effective than not providing them, thus the condition without transparency was excluded.

2.1 HAT Platform

2.1.1 Game Task

A Pictionary game in which a person creates a drawing, and an AI guesses the image has been deployed on several online platforms. Among them, ‘QuickDraw’ developed by Google is the one most popular and most accessible. QuickDraws is an online platform that challenges users to draw pictures of everyday items and predict what the drawings represent with a deep neural network (DNN) algorithm (Cai et al., 2019). It has been used by millions of people worldwide, thus there is large amount of accumulated data drawn by diverse people. Therefore, it was selected to design the experimental simulator for this study by mimicking the game method and interface of QuickDraw.



What do 50 million drawings look like?

Over 15 million players have contributed millions of drawings playing Quick, Draw! These doodles are a unique data set that can help developers train new neural networks, help researchers see patterns in how people around the world draw, and help artists create things we haven't begun to think of. That's why we're open-sourcing them, for anyone to play with.

Figure 2. QuickDraw online platform.

2.1.2 Simulator

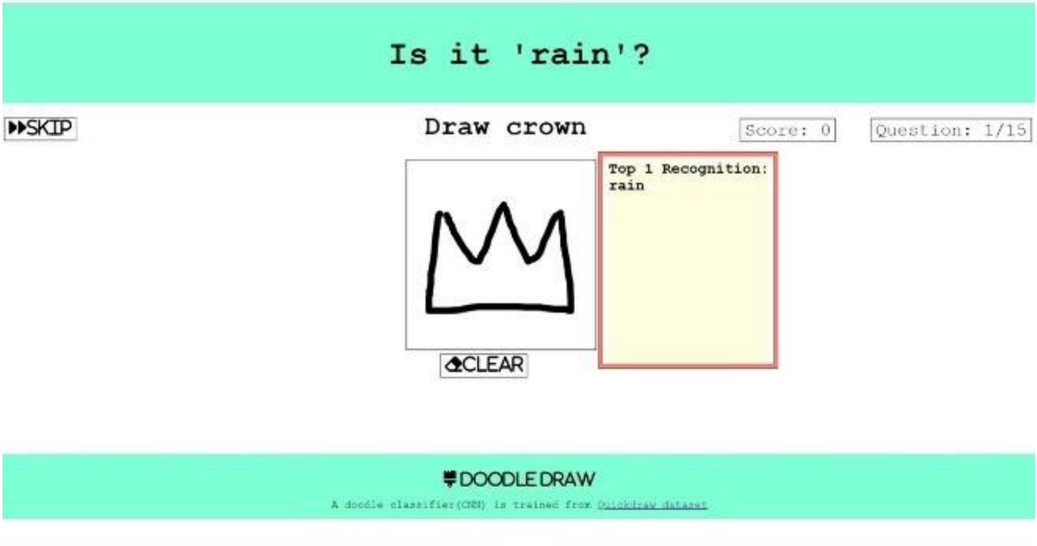
The experimental simulator was built with the open-source AI engine from GitHub. Yining and Zhiwei (2018) developed a doodle classifier with Convolutional Neural Network (CNN) model using python and a website showing what image people are drawing in real time using this classifier with HTML and JavaScript. CNN models are particularly useful for recognizing objects, faces, and other patterns in images by finding specific patterns in the image data. This doodle classifier utilized a dataset released by QuickDraw, consisting of a total of 345 categories, with each category containing 50,000 data used for training and testing. The model achieved an accuracy of 89.67%. Although attempts were made to enhance the performance of the AI engine, the vast amount of data proved to be too much for the computer's capabilities, thus AI engine was used as is. Since the website's interface and task needed to be redesigned to perform the experimental task, the existing HTML and JavaScript were modified, and CSS codes were added.

The final prototype of the simulator was developed under the name "Doodle Draw," as shown in Figures 2 and 3. There was a communication panel at the top where AI guesses and checks whether the object it guessed is correct. Through this communication panel, the AI was able to engage in two-way communication with humans in natural language. The AI continuously made sentence-level predictions about what the given word could be. If the AI correctly identified the image, it would display "Correct! Great Job." However, if the participant skipped the question since the AI was unable to correctly identify it, it would display "Sorry, I couldn't get it".

Right below that, from left to right, the skip button, the proposed word for the player to draw, the current score, and the number of current questions were listed. In the center of the screen, there was a canvas where the participant can draw, with a clear button located directly below for erasing. The core of this interface was the transparency information, which is displayed in a yellow box highlighted in red next to the canvas. The top recognition information showed the top few items that the AI system recognizes as being most similar to the participant's drawing. The simulator with low transparency displayed only the top 1 object predicted by the AI, whereas the simulator with high transparency showed the top 5 objects predicted by the AI, along with the predicted percentage for each. In this way, the number of cues predicted by the AI has been controlled to display either 1 or 5.

2.2 Task

A set of 36 of the most common everyday words with high recognition rates were extracted and divided into two practice trials of 3 words each and two test trials of 15 words each. When the experiment began, one of the 15 assigned words was randomly presented on the screen. After the participant drew the word, the AI continued to make predictions in real-time, and the system presented a list of 1 or 5 items in the top recognition information section that AI predicts to be the most similar. Participants modified their drawing while looking at the words predicted by the AI, and when the AI correctly identified the word, one question was passed. This repeated until a total of 15 questions were fully completed.


Is it 'rain'?

Draw crown

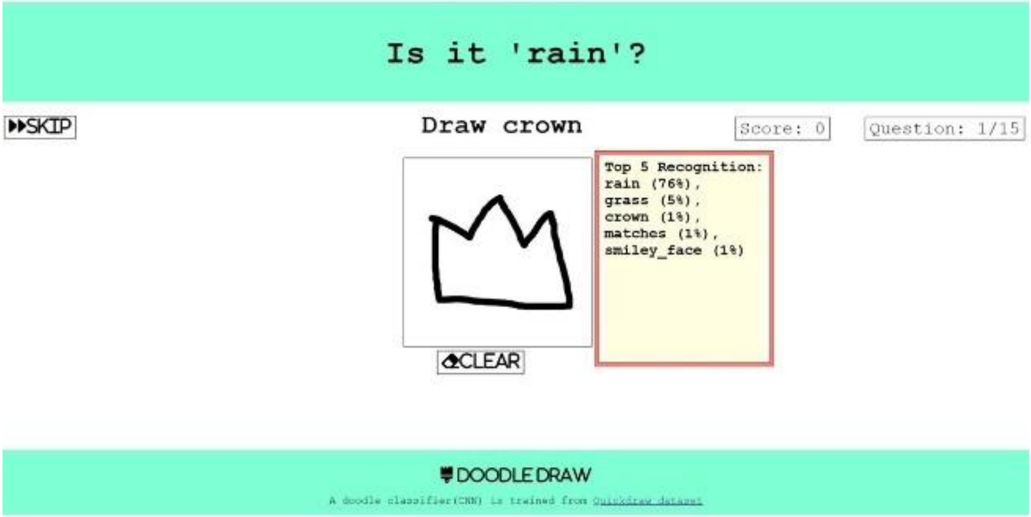
Score: 0

Question: 1/15

Top 1 Recognition:
rain

A doodle classifier(CNN) is trained from Quickdraw dataset

Figure 3. Simulator with lower transparency showing top 1 recognition.



Is it 'rain'?

Draw crown

Top 5 Recognition:
rain (76%),
grass (5%),
crown (1%),
matches (1%),
smiley_face (1%)

A doodle classifier(CNN) is trained from Quickdraw dataset

Figure 4. Simulator with higher transparency showing top 5 recognitions.

Table 1. Word set for each practice and test trials.

Trial (# of words)	Word Set
Practice A (3)	‘ocean’ ‘fork’ ‘bench’
Practice B (3)	‘sun’ ‘smiley face’ ‘knee’
Test A (15)	'crown', 'snowflake', 'laptop', 'stairs', 'cactus', 'remote control', 'diamond', 'hamburger', 'pillow', 't-shirt', 'light bulb', 'hammer', 'mug', 'camera', 'hurricane'
Test B (15)	'necklace', 'traffic light', 'jail', 'book', 'mailbox', 'alarm clock', 'door', 'candle', 'ladder', 'pool', 'microphone', 'grass', 'strawberry', 'chair', 'drill'

2.3 Subjects

A total of 20 participants were recruited for the experiment. Participants were required to be fluent in English and have no specific constraints on physical and cognitive abilities for gaming. The age range of participants is 20-35 years old, with a mean of 26.55 years old. Eleven were female and nine were male. All participants participated voluntarily.

Prior to the practice trial, participants were requested to take a survey as part of the experimental procedures in order to identify the general characteristics of the sample population.

They were asked how often they sketched on a computer and how good their drawing skills were. Subjects reported an average rating of 2.1 (sd = 1.01) for the frequency of drawing on a computer on a scale ranging from 1 (= Never) to 5 (= Daily). Participants gave an average rating of 2.1 for their level of drawing skill (sd = 0.83) on a scale of 1 (=Beginner) to 5 (=Proficient). This indicated that the participants drawing with a computer were relatively inexperienced and correspond to the elementary level.

All experimental procedures, including recruiting, data collection, and data analysis, were conducted after obtaining IRB exemption from the Institutional Review Board (IRB) at the University of Michigan after review of the research procedure (HUM00224796).

2.4 Experimental Design and Variables

The experiment followed a within-subject design. Participants completed one test trial with each level of transparency. A total of 40 trials (20 participants x 2 trials) were completed along with the associated data set. The order of presenting level of transparency, word set, and words within a word set was randomized across participants and trials to prevent an ordering effect. The key independent variable (IV) of the experiment was the two different levels of transparency (Top 1 object/Top 5 objects). The total of 30 words provided in two test trials is another independent variable.

Dependent variables (DVs) were broadly divided into two types based on whether they were word-by-word or condition-by-condition. First, word-by-word dependent variables were time-to-task completion (TTC), number of errors, and eye profile data. TTC was measured as the time to complete each word. Participants were told to complete the test as quickly as possible. The number of errors referred to the number of times the clear button was clicked to clear the image per word. The eye profile data was measured in two aspects: dwell time which was the amount of times (sec) that the eyes remained on the transparency information screen while performing a single word, and the frequency in which the eyes moved to the transparency information screen while executing a single word.

Dependent variables by the condition of the transparency level were performance, the total number of errors, subjective workload, and subjective trust. In each trial, score was calculated by multiplying the total number of words matched by ten. The total number of errors referred to the number of times clicking the clear button to clear the image by the condition of the transparency level. The NASA-TLX scale was used to measure subjective workload in each test trial. Subjects completed the rating form after each test trial, and the demand ranking form was completed once all test trials were completed. They were also asked to assess the subjective trust of each system after each test trial. Along with the NASA-TLX questionnaire, participants were also asked to rate their subjective trust in each system. The subjective trust survey was completed by subjects after each test trial. It consisted of six 7-point Likert scale questions with different dimensions of trust. The question assessed system **understanding** (“I understand what the system is thinking”), **capability** and **benevolence** (“The system seems capable”, “The system seems benevolent”), **ability** (“The system has the ability to deliver knowledgeable information”), as well as **integrity** and **reliability** (“The system seems to be integrated and reliable”), key dimensions of trust that was widely used in trust questionnaires (Cai et al., 2019; Mayer et al., 1995; Seppänen et al., 2007; Khalid et al., 2016). **Attribution** of the recognition result (1 = “Totally due to the recognizability of my drawing”, 7 = “Totally due to the system’s level of capability”) was also included in the survey (Cai et al., 2019).

A post-trial interview was done following the experiment to collect human factors issues that occurred during human-AI teaming. Initially, the individuals were asked whether they preferred a system with top 1 object or top 5 objects, and the reasons for their preference were investigated. They were then asked whether having more of AI's top recognition information enhanced their understanding of AI or offered guidance on how to draw better. The participants were then asked why they think AI identified their drawings well or poorly depending on their performance. Lastly, general feedback on the experiment was requested, and the experiment was complete.

2.5 Procedure

Each participant was required to complete the whole experiment using the following procedures: (1) They spent 10 minutes on administrative details and a description of the process, task, and equipment for the experiment. They also completed paper documentation, including a consent form and a demographic questionnaire. (2) They were given two practice trials with instructions for each level of transparency to become familiar with the drawing task incorporating AI. They drew three items provided in each trial until the AI correctly answered. (3) They performed two test trials with instructions. Before starting, they were asked to get the correct answer as many of them as possible. After completing each test trial, the participants finished a NASA-TLX workload rating survey and a subjective trust survey. (4) After completing all test trials, the participants were asked to complete a NASA-TLX workload ranking survey and express their general comments about the experiment. The entire experiment lasted about 60 minutes, with breaks provided whenever the individual requested them.

2.6 Hypothesis & Data Analysis Plan

The top recognition information aimed to promote transparency about what AI thinks. It was hypothesized that if the top 5 AI recognitions were presented rather than just the top 1 recognition, it would improve the participants' task performance. Furthermore, it was predicted that increased transparency would result in a higher workload for participants. Finally, it was anticipated that when the top 5 AI recognitions were displayed, participants would gain a better understanding and trust in AI.

For quantitative analysis, ANOVA tests were conducted on the task performance and eye profile data as well as the subjective workload and trust ratings to determine whether there are significant differences between the two levels of transparency. Additionally, for qualitative analysis, comments collected through post-trial interviews were analyzed using open-coding to identify critical comments and examine important human factors issues in human-AI teaming.

CHAPTER 3: RESULTS AND DISCUSSION

3.1 Time to Task Completion (TTC)

To address individual differences between participants, all TCC data was standardized (converted to z-scores). Data were standardized by subtracting the mean value from each data point and dividing it by the standard deviation. This procedure uniformed the data to a consistent scale, allowing it comparable and enabling outlier handling. Analysis of variance (ANOVA) results showed that the mean task completion time for the two levels of transparency was not significantly different ($F_{1,512} = 2.07$, $p = 0.151$). Transparency had no apparent impact on task completion time for participants. However, as seen in figure 4, the task completion time of participants rose considerably when transparency was increased. This appeared to be due to the increased amount of information that participants need to process as transparency increases.

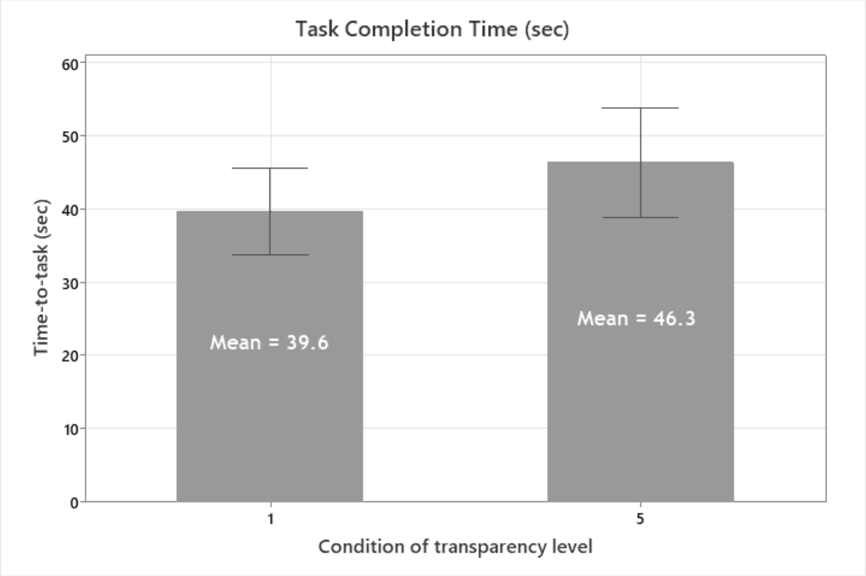


Figure 5. Task completion time (in sec) for two different transparency levels.

3.2 Number of Errors

Results of the non-parametric (Kruskal-Wallis) test for the effect of transparency level on the number of errors per word (χ^2_1=0.00, p = 0.975) and number of total errors per trial (χ^2_1=2.63, p = 0.105) were not significant. There was no significant difference on the number of errors between the two levels of transparency, either by word or by trial. Even though the system provided more top recognition information to participants, it did not appear to enable participants to make fewer changes to their drawings in order for AI to correctly guess the answer.

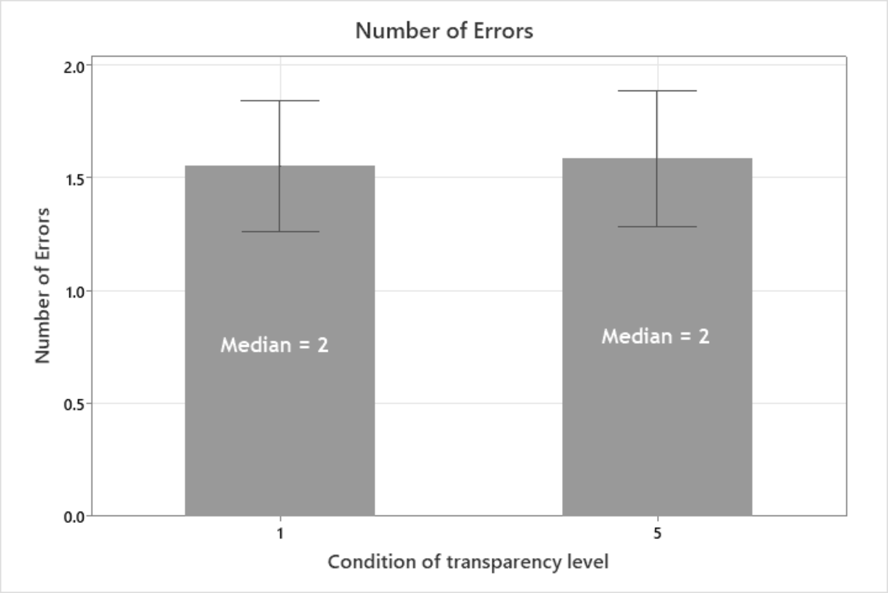


Figure 6. Number of errors per word in two different transparency levels.

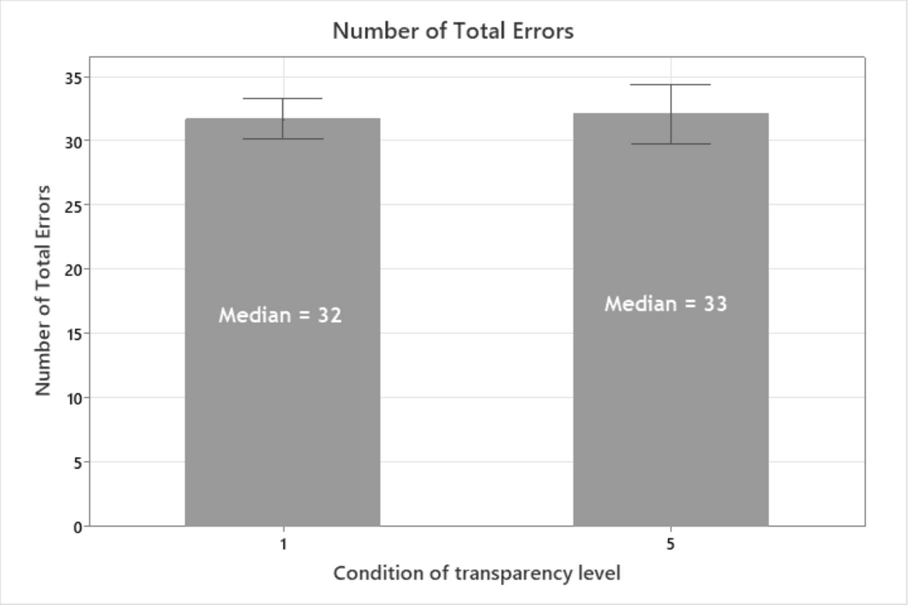


Figure 7. Number of total errors per trial in two different transparency levels.

3.3 Eye Profile data

The dwell time and frequency of eye movement were gathered by manually monitoring the participants' eye movements in recorded videos of their faces during the experiment, using Adobe Premiere to track pupil movements. Both dwell time and frequency of eye profile data were standardized (z-scores) to address individual differences such as eye movement speed. For eye dwell time, analysis of variance (ANOVA) was conducted, and for frequency of eye movement, the non-parametric Kruskal test was performed. The results revealed that only the mean dwell time ($F_{1,512} = 38.7$, $p < 0.001$) had significantly difference between the two levels of transparency. Frequency of eye movement didn’t not have significant difference for two levels of transparency ($\chi^2_1 = 1.79$, $p = 0.18$). When presenting AI’s top 5 objects instead of top 1, it was observed that the participants’ eye dwell time rose as the amount of the information they needed to read increased. This suggests that as the transparency increased, participants tended to refer to the additional information for a longer period.

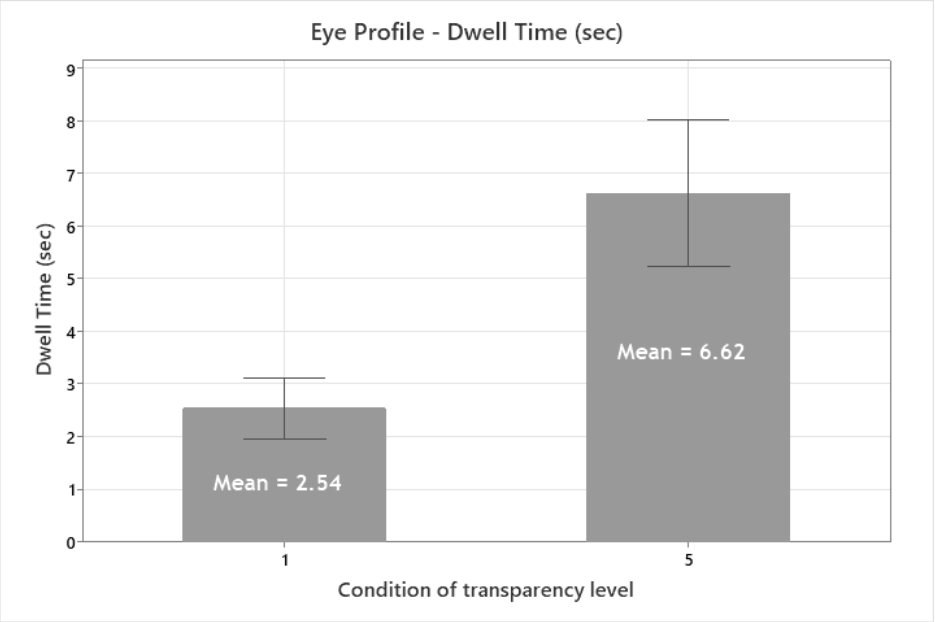


Figure 8. Dwell time (sec) per word in two different transparency levels.

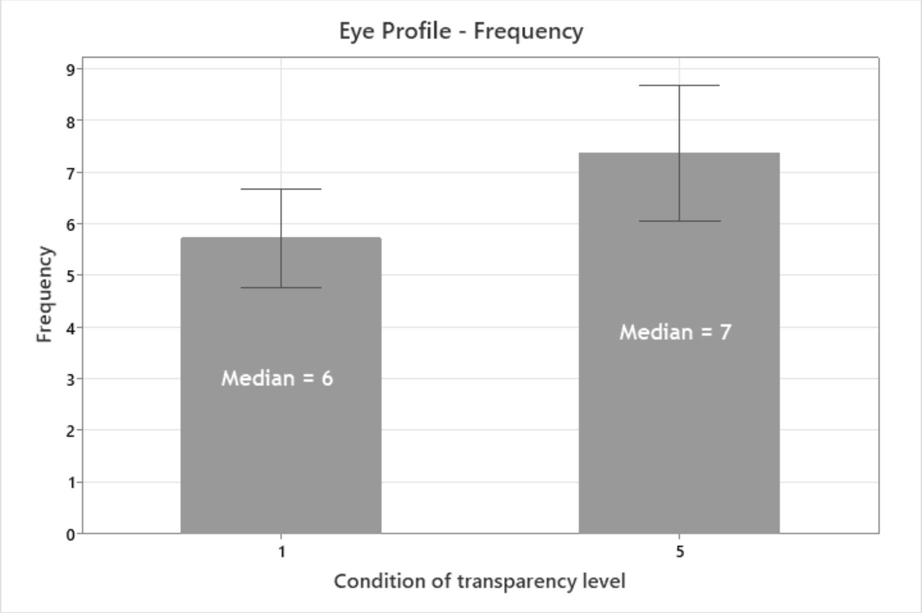


Figure 9. Frequency per word in two different transparency levels.

3.4 Task Performance

Result of the non-parametric (Kruskal-Willis) test for the effect of transparency level on task performance was not significant ($\chi^2{}_1$=0.52, p = 0.72). There was no significant difference in scores between the two levels of transparency. Even though AI provided more transparent information to participants, the AI did not appear to be able to match more drawings. There could be multiple reasons for this finding, but two main causes were suggested. One possible explanation is due to the absence of time pressure. Since there was no time limit during the experiment, the participants made multiple modifications to their drawings to receive as many points as possible. Another possibility is that even in higher-transparency condition, there might have been limitations for AI to accurately identify the correct answers, no matter how much the participants challenged themselves. However, further analysis is needed.

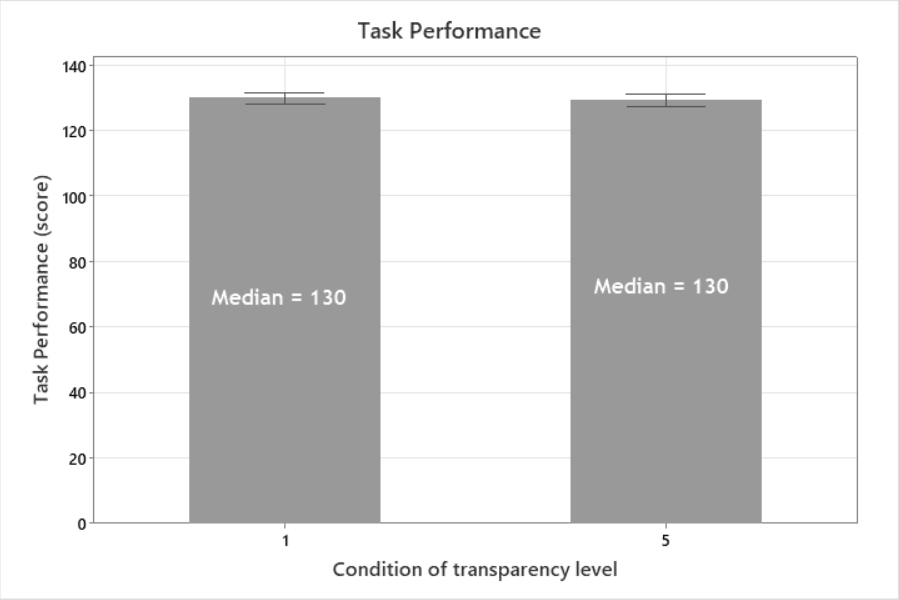


Figure 10. Task Performance in two different transparency levels.

3.5 Subjective Workload

An ANOVA was performed on subjective workload scores. Result revealed that mean of overall workload ratings for two transparency level to be significantly different ($F_{1,512} = 35.73$, $p < 0.001$). As shown in Figure 10, it was observed that the overall subjective workload decreased significantly when the number of top recognition information increased to 5.

A series of ANOVA tests on ratings for each dimension in NASA-TLX provided a more detailed explanation for differences in workload. Among the 6 dimensions of NASA-TLX (mental demand, physical demand, temporal demand, performance, effort, and frustration), the ratings on mental demand ($F_{1,512} = 8.38$, $p = 0.004$), physical demand ($F_{1,512} = 10.97$, $p = 0.001$), temporal demand ($F_{1,512} = 20.27$, $p < 0.001$), and frustration ($F_{1,512} = 20.64$, $p < 0.001$) were significantly different for the two transparency level. Figure 11 illustrates that all four dimensions of subjective workload showed a substantial decrease when the higher-transparency condition was presented. This finding demonstrated that increased transparency had the effect of reducing workload, which was a different tendency from previous studies (Kunze et al., 2014a; Chen et al., 2018; Yang et al., 2017; Wright et al., 2019; Mercado et al., 2016).

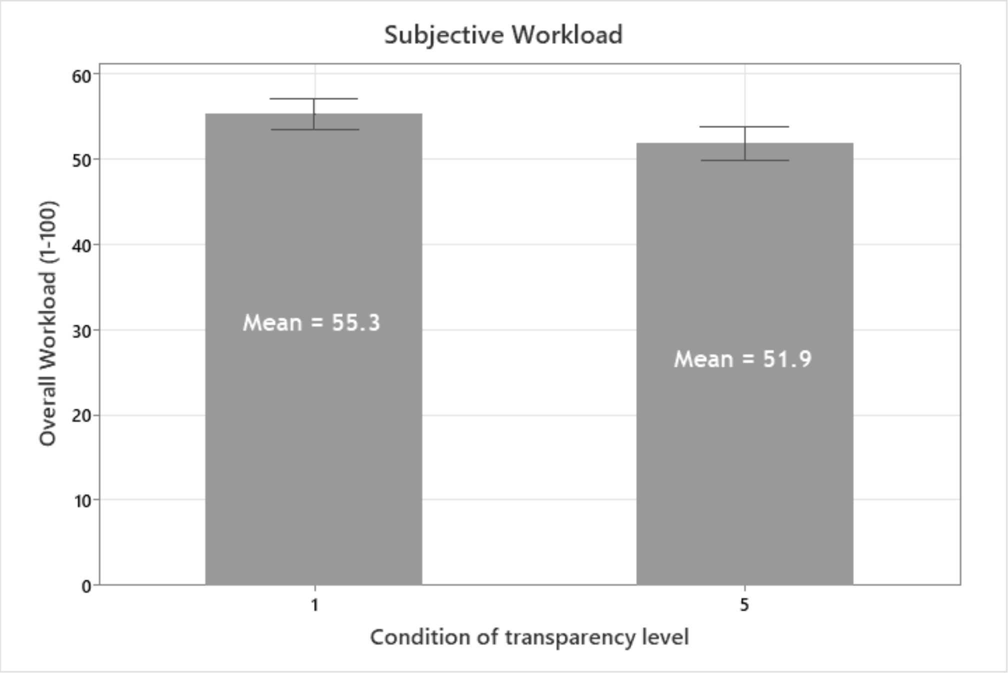


Figure 11. Overall workload in two different transparency levels.

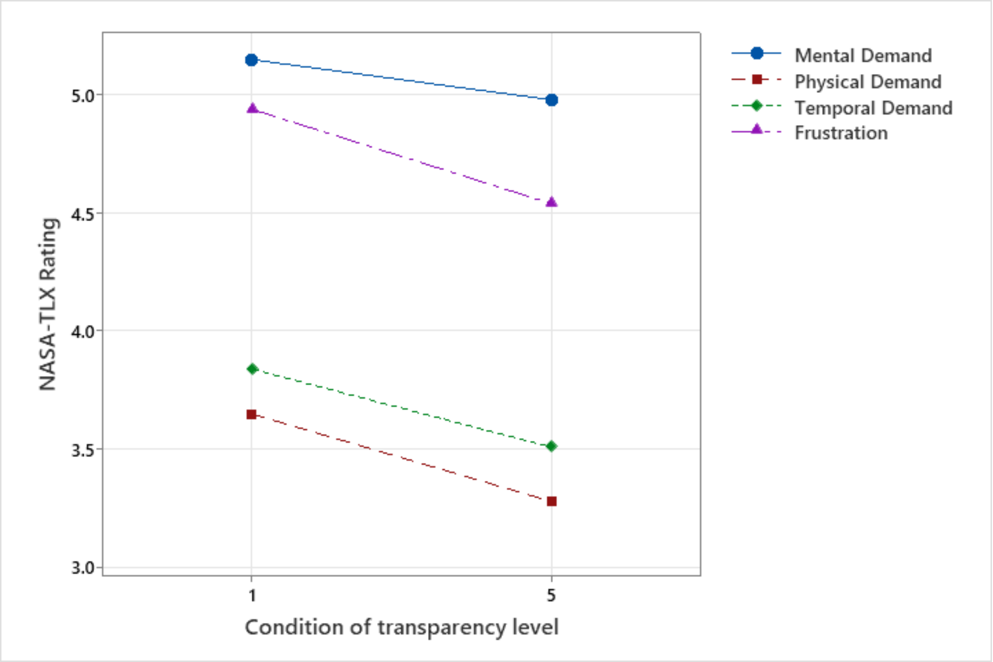


Figure 12. Four dimensions of subjective workload in two different transparency levels.

3.6 Subjective Trust

An ANOVA test was conducted on the six aspects of subjective trust in the survey. The mean rating of five aspects, which were understanding ($F_{1,512} = 12.49$, $p < 0.001$), capability ($F_{1,512} = 33.43$, $p < 0.001$), benevolence ($F_{1,512} = 4.06$, $p = 0.045$), ability ($F_{1,512} = 32.17$, $p < 0.001$), and attribution of the recognition result ($F_{1,512} = 1.40$, $p = 0.237$), was significantly different for the two different transparency levels. Interestingly, these five dimensions were divided into two groups and had opposing patterns. Some aspects showed higher mean ratings in the higher-transparency condition, while others received higher mean ratings in the lower-transparency condition.

It was discovered that the participants' understanding of AI and trust in its capabilities were substantially higher when the AI showed the top 5 similar objects. Mean ratings for the questions "I understand what the system is thinking" ($M = 4.67$ on the top 1 object, $M = 4.94$ on the top 5 object) and "The system seems capable" ($M = 4.82$ on the top 1 object, $M = 5.1$ on the top 5 objects) were higher for the system with higher-transparency condition. A post-trial interview revealed that displaying AI's top 5 objects informed participants on what the drawing was similar to and what features it was catching. The participants stated that their understanding of AI improved when the top 5 recognition information, along with the corresponding similarity percentages, were shown on the system. The participants claimed that the system displaying the percentage along with the top 5 objects demonstrated higher capability compared to simply showing the top 1 object.

On the other hand, questions about benevolence and ability had opposing outcomes. The system with a lower level of transparency received higher ratings for the questions "The system seems benevolent" ($M = 4.6$ on the top 1 object, $M = 4.5$ on the top 5 objects) and "The system has ability to deliver knowledgeable information" (M = 5.1 on the top 1 object, M = 4.88 on the top 5 objects). This outcome is assumed to be the consequence of two reasons. First, when showing the top 5 objects, participants showed a tendency to focus on drawing features that would enable AI to recognize the object correctly, rather than creating subjective images of the objects. However, when providing only the top 1 object information, it was difficult to know what features the AI was catching, so the participants drew the images based on their personal creativity and subjectivity. Therefore, they might have felt that the system with lower-transparency condition was more benevolent because it recognized the image they drew based on their subjective opinions. In post-trial interviews, some participants mentioned that although the transparency of the AI increased, they felt frustrated and faced a great sense of disappointment when the system showed items that were completely unexpected. They also mentioned that more forgiveness was required for the system with a higher transparency level.

Another possibility is as follows. When participants used systems with two different levels of transparency, a different tendency was discovered in the areas where their pupils lingered. When showing the top 5 objects, participants focused on viewing the top five similar items and corresponding similarity percentages. On the other hand, when showing the top 1 object, they were found to pay more attention to the communication panel with AI at the top. The communication panel displayed an AI-generated phrase predicting the most similar object. It also apologized to participants if they skipped a task that AI couldn't complete correctly. Therefore, participants might have felt that the system showing top 1 recognition information conveyed more benevolent and knowledgeable information. However, further analysis is required.

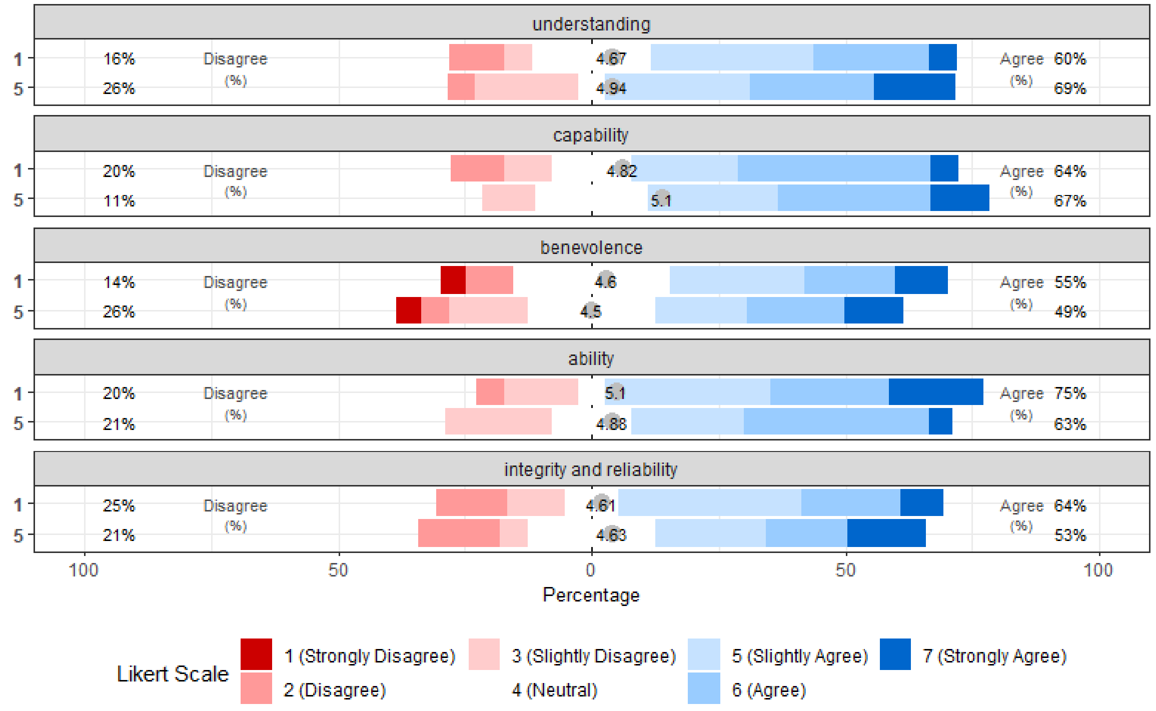


Figure 13. Subjective trust in two different transparency levels (Dimension of understanding, capability, benevolence, ability, integrity and reliability).

The mean rating for the question "The system appears to be integrated and dependable" did not demonstrate a significant difference between the two levels of transparency ($F_{1,512} = 1.4$, $p = 0.237$). Lastly, the study found a significant difference in the mean rating for the attribution of recognition results under the two transparency conditions. As illustrated in Figure 13, the participants indicated that the system's capability, rather than their own efforts, was responsible for the drawing's recognition when the transparency level was low. On the other side, when the transparency level was high, they felt that their own efforts were the reason behind the image's recognition. When the AI transparency was high, the participants were able to figure out what features the AI was capturing through the top 5 recognition information and corresponding similarity percentages. Thus, they reflected those features in their drawing to enable the AI to correctly recognize it. In contrast, with a lower-transparency condition, the participants had no information about what features the AI was catching, so they had to draw the image based on their assumptions. In other words, they drew images based on their subjective opinions until the AI recognized the picture correctly. Therefore, while higher transparency improved the participants' understanding of the AI itself, it required more cognitive and physical efforts for them to make the image recognizable. Conversely, when the transparency was lower, the participants drew images based on their subjectivity, and AI had to match the images through its own image recognition ability.

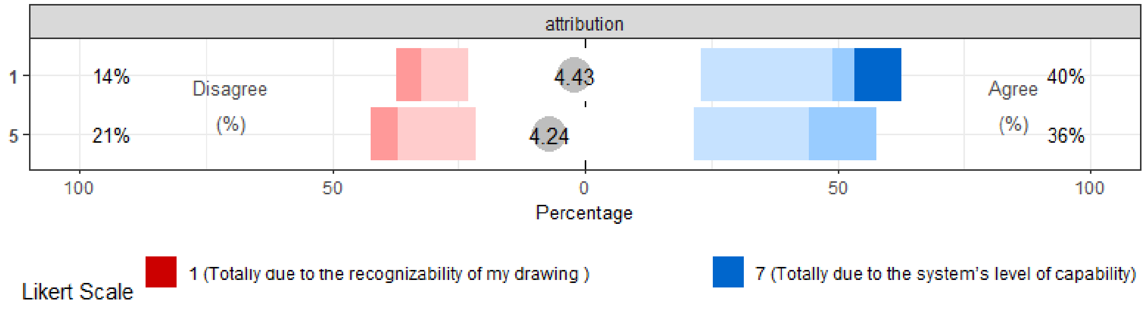


Figure 14. Subjective trust in two different transparency levels (Dimension of attribution of recognition results).

3.7 Post-Experiment Questionnaire and Interview

After the experiment, post-experimental interviews were conducted with all 20 subjects. To discover some critical human factors issues during experiment, sentence-by-sentence open coding was conducted.

When participants asked whether they preferred a system with top one or top five recognition information, 16 participants selected the top five recognition system, while the remaining four preferred the top one recognition system. There were three primary reasons why participants preferred the system with higher transparency. Firstly, they used the similarity percentage that accompanied the top five objects as a proximity indicator to determine how close or faraway the target was. By tracking the changes in top recognition information and similarity percentages each time they modified the image, it was feasible to assess whether they were getting closer or further away from the correct answer. Another reason why they preferred the higher-transparency condition was that they felt an increased sense of mutual understanding between themselves and the AI. Participants stated that the system showing top 5 objects was more likely to understand their drawings accurately and provided the understandings of AI's decision. This mutual understanding was seen as crucial to individual preferences. Furthermore, participants indicated that greater transparency of AI improved their understanding of AI's decision and allowed them to see that the AI could identify what they expected.

On the other hand, those who preferred top 1 recognition had the following reasons. First, participants experienced confusion and frustration when testing the system with top 5 recognition, especially when the suggestions provided by the AI were unexpected. There was a mismatch between them and the AI in comprehending the images for the same object. Therefore, they felt more at ease when using top 1 recognition. Several participants preferred the system with binary response, showing whether or not it correctly matched the target, because it was more intuitive and easier to interpret. They also preferred the system with top 1 recognition since the system with higher transparency provided too much information which increased their workloads.

Table 2. Critical comments on preference for two different level transparency

Codes	Participants (#)	Critical Comments
Preferred Top 5 Recognition		
Proximity indicator	Participant 2	“That give me a better sense of what I might be doing wrong, cause the percentage kind of keeps telling you how closer or how further away you are from the actual target.”
	Participant 3	“I was just looking at the number, then the percentage of strawberry increased. I’m like okay, the system requirement is this. I tried to relate what I had drawn.”

	Participant 20	“I could see whether my drawing is getting closer because of the percentage”
Mutual understanding	Participant 10	“That taught me a little bit more about what I was expecting, and just kind of what it would be like.”
Transparency in AI reasoning	Participant 6	“Because with top 5, I can manage to make the AI system understand what I’m doing.”
	Participant 13	“By looking at what shapes are really close to the shape you're drawing, you're getting the better insight from what's the system of understanding.”
	Participant 18	“Thinking of the AI becomes a little bit more transparent for me.”
Preferred Top 1 Recognition		
Mismatch of understanding with top 5 recognition	Participant 4	“I found like when I got the top 5 of how it was thinking, it made it a little harder for me to get the address. It made it harder for me to address what I was thinking, top 1 recognition helped me stay more creative, and finding other ways to explain to the AI. Where, as when it looked at the top 5 and mine was not one of them, no matter what I'm gonna do, I'm never getting. I got a bit more frustrated.”
	Participant 11	“When with the top 5, if it gives you a list, and then there may be the one that you're looking for, and you're trying to improve your diagram, the more you try to be the further away it goes from the one and I’m like okay. Why was I even trying?”
Intuitive understanding	Participant 15	“I see my drawing is somewhere here, so I’m trying to make it look similar. But as that one was just it's either one or 0. It's binary. and I think that is more my side.”
Less mental workload	Participant 19	“Top 1, it required less mental workload. If I have more for it, I may have made it the thinking more complicated.”

Then, the subjects were asked whether the increased transparency had improved their understanding of AI. Most of the participants, 18 out of 20, replied that the increased transparency raised their understanding of AI. The remaining two stated that they had not gained the understanding of AI because there was a significant gap between the image that they had in mind and the image that the AI had in mind for the same object.

When the participants were asked whether increasing transparency provided them more knowledge on how to draw, they had different point of views. There were 5 participants who strongly agreed, 10 who somewhat agreed or disagreed, and 5 who completely disagreed. Those who agreed were able to identify what details to include in the image by capturing the features that the AI was looking for. For individuals who disagreed perceived a big difference between the image they had in mind and the image that the AI had in mind for the same object. Therefore, they responded that they had no idea how to draw the image appropriately. Several participants offered some critical feedback. They stated that additional information was required to understand why their drawing was perceived to be similar to the AI's suggestions. This emphasizes the need of explaining the mechanics and procedures by which the AI identifies similarities between their drawing and other items.

According to the National Academies of Sciences, Engineering, and Medicine (2021), transparency includes providing information on the purpose, progress, and performance of AI systems, which encompasses an understanding of the current state of the system. The system's behavior and recommendations should be understandable to human teammates, including not just the system's functionality but also the logic or factors driving its behavior. Thus, as evident from the feedback of the participants, it was revealed that there was a lack of information regarding the system's decision-making process and underlying reasons. Future research should expand transparency information to include the information about AI's progress and reasons.

Table 3. Critical comments on how to improve transparency of AI.

Codes	Participants (#)	Critical Comments
Need additional explanation	Participant 3	“It was not to cheat the system. How does this look like 3% of it? How and why? But I know this system recognized this. I didn’t get what’s the how and why.”

Next, participants were asked to provide their subjective opinions on the key factor influencing AI image recognition success or failure. Their perspectives were divided into three categories: AI feature detection, conceptual difference or similarity of object imagery, and the need for additional explanation or support.

One of the most critical aspects was whether participants could identify the features that AI detected in the drawings. As previously stated, the participants tended to draw their images in a manner that would enable the AI system to recognize correctly. Considering the characteristics of CNN models, which quickly capture the features of images, it appeared that people noticed this characteristic throughout the experiment. Participants tried to discover which specific parts of their drawings were being detected by the AI and adjusted their drawings based on the transparency information.

Another critical factor mentioned by a significant number of participants was the conceptual difference or similarity of object imagery. AI learned typical images of each object through its training data. In order to obtain the correct answer, it was crucial for the learned object imagery of AI to correspond with the mental images that participants had for each object. Therefore, if the participants had different object imagery, they made several attempts to match the object imagery of AI. However, what AI recognized from its point of view was very limited. For example, as seen in figure 14, the form of a camera has gradually changed as generations have passed. Film cameras were used in the early 1900s, and digital cameras were popular in the late 1900s. Nowadays, most people primarily use cameras embedded into their smartphones. Unfortunately, the AI recognized only the images of digital cameras. This demonstrated AI's lack of understanding in learning individual differences such as generation gaps. Therefore, as mentioned by participant #18 during the interview, it appeared that images of objects drawn from various perspectives should be included in the training data. This refers to not only images of objects drawn from various angles, but also images of objects that contain individual differences such as cultural backgrounds, generational gaps, and gender differences. Several studies revealed that people are more likely to trust and work with autonomous agents that exhibit similar work style preferences to their own (You & Robert, 2018). The finding may be tied to the similarity-attraction effect (Byrne, 1997), but more studies are needed to confirm its robustness. Thus, it will be crucial to investigate how incorporating individual difference learning into AI affects human factors like trust and team performance.

Lastly, it was indicated that one of the reasons why the AI failed to correctly identify the drawings was due to a lack of explanatory information. The type of information that can enhance explanatory power varies, but throughout the interview, it was clear that additional transparency in AI's decision-making process was needed. This falls under the category of 'explainability,' which reveals the logic behind AI's decision-making process. It demonstrates that displaying transparency alone was insufficient for participants to improve the success rate of AI image recognition. Additional explanations were necessary for participants to understand which features the AI recognized, how it derived the top recognition information and corresponding similarity percentage, and how the AI correctly or incorrectly identified the drawings. Therefore, it was determined that displaying transparency alone has limitations, and explainability should also be provided to improve the success rate of AI image recognition.

Table 4. Critical comments on the key factors behind the success/failure of AI's image recognition.

Codes	Participants (#)	Critical Comments
AI feature detection	Participant 5	"I think it's because I knew I adjusted to what the AI look for."
	Participant 16	"The reason why is that it needs specific details that it will give you the percentage changes. It gives the same percent if I just

		keep drawing the same thing again. I can see the same thing, pattern.”
Conceptual difference/similarity in object imagery	Participant 2	“I think it would be a perception thing apart from my drawing, my imagination skills. Or maybe the software isn't imaging what I'm imaging.”
	Participant 4	“I felt like there was a bigger gap than I would have thought. I'm between my drawing and what it was thinking.”
	Participant 13	“I was trying to sketch it from one perspective, and it always keeps failed to recognize my drawing, and then when I get through it from the other perspective, it got successful.”
	Participant 18	"Training data needs to be bigger be good and including different, like diverse perspectives, and that includes like different gender ethnicity, different cultural groups. It's really important to capture all this perspective when you are aiming to have one universal system, AI system, to recognize everything it need to start learning cultures, different perspectives, age groups.”
Additional explanation/support needed	Participant 9	“If you really want to see how people are drawing, and then how the system is understanding, and you can provide a reference images.”
	Participant 18	“It's still unclear at what detail it starts recognizing as it. It was another kind of a big question mark like, how much detail does it need to recognize?”
	Participant 19	“If you're testing a system, one like, user begin with your use this system, you can provide some good examples and pictures, or to to start with it.”

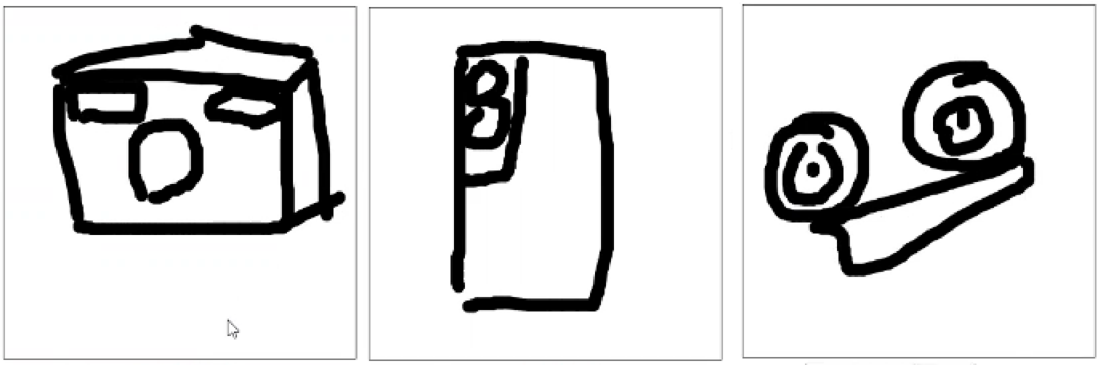


Figure 15. Example drawings of camera.

Participants were requested to provide general comments at the end of the experiment, which was categorized into three different aspects. Firstly, several participants noted that the AI required additional training. While some respondents stated that the AI's reliability and performance were adequate, others stated that the AI failed to recognize some items. Moreover, there were frequent errors with the AI identifying same drawings as completely different objects. Therefore, participants suggested that additional training was needed to improve the performance of the AI’s recognition.

Another feedback was that there were several words that had the similar appearance in the 345 categories, requiring a categorization process. For example, when participants were given the term "mug" to draw, the AI recognized their drawing as 'cup' or 'coffee cup' instead, causing confusion to the participants. Therefore, there was a need to integrate words with similar shape.

There were also recommendations for modifying the experimental design or enhancing the experimental tool. The experiment in this study was designed to match as many given items as possible without a time limit, and participants were instructed to create drawings on a computer-based simulator with a mouse. One suggestion was to include a time limit to generate time pressure during the experiment. The other idea was to allow participants to draw directly on a tablet with some painting tool to change line thickness and color.

3.8 General Discussion

In general, a system that displayed the top 5 objects with a higher level of transparency outperformed the other system except for task completion time, number of errors, task performance, and some dimensions of trust. It was mainly due to the AI's performance issues and the need for additional explanation or support. The AI had an accuracy rate of 89%, but during the experiment, it showed significantly lower recognition rates for some items. It highlighted the need to add diverse images of objects from various angles and individual differences such as cultural difference and generation gap. It also pointed out the need to add additional layers to the CNN model to improve performance when training the AI. Moreover, AI was trained to classify drawings into 345 categories, but similar-shaped items were not grouped, which confused the participants. Therefore, additional categorization tasks were required to group similar objects.

Although the top 5 objects that AI considers most similar to the drawings were provided with corresponding similarity percentages, more detailed explanation was required. Explanatory power could be provided by showing explanations on which feature of the drawing that AI caught to determine the top recognitions, and information on how or why the AI successfully matched or failed to match the drawings. It was expected that offering such explanations would assist participants better understand the AI and how to draw the images. In conclusion, it appeared that display transparency alone had limitations in increasing human trust and team performance. Therefore, the need for providing explainability along with display transparency had been identified.

CHAPTER 4: CONCLUSION

This study compared two different levels of transparency in a Pictionary game system that demonstrated top 1 object and top 5 objects of AI. The experiment was conducted to investigate how increased transparency affects performance and human trust. The results showed that increased transparency did not affect task completion time, number of errors, and task performance. Contrary to hypothesis 1, transparency of AI did not improve team performance under the experimental conditions. However, displaying higher-transparency condition significantly reduced subjective workload. Increased transparency had a beneficial effect on subjective effort decrease, which was completely opposite to hypothesis 2 in this study. Lastly, increased participants' understanding of AI and their trust in the system's capability. Yet since some dimensions of trust declined as transparency increased, hypothesis 3 was partially proven.

This study identified additional important human factors issues that arise in human-AI teaming. Through post-trial interviews, it was found that mutual understanding between participants and AI is a crucial factor that greatly affects the experiment results. Additionally, it was observed that an AI system that considers individual differences has the potential to dramatically improve team performance and human trust. Yet, this issue has not been adequately addressed in previous studies. Therefore, this study implies that further investigation of this subject is required in the future. Lastly, it was discovered that display transparency alone has limits in understanding the logic and decision-making processes of AI systems. Thus, explainability should be offered in addition to transparency to attain its full effect.

However, there were some caveats to this study. Firstly, differences in performance were observed depending on which system participants experienced first. While there was a practice trial for each system, it is possible that the time provided was insufficient for participants to become fully familiar with the system. Thus, few individuals showed better performance on the second test trial as they gained a better understanding of the AI system. This might have led to unfair comparisons between the two different levels of transparency. The second limitation was that there was insufficient balancing in difficulty when separating two different word sets. Therefore, it might have had a negative influence on comparing the effects of two different levels of transparency. It not only implies a need for higher accuracy in AI recognition for each word, but also emphasizes the significance of extracting and distributing words into word sets of similar difficulty. The final limitation was the insufficient number of transparency levels. To discover an optimal level of transparency that exposes the most beneficial outcomes, transparency must be divided into more diverse levels. However, this study only evaluated only two levels of transparency.

In regard to future research, it would be interesting to divide transparency levels into more detailed and investigate how much transparency would be most beneficial. Furthermore, since display transparency alone had limitations in increasing human trust and team performance, it would be interesting to provide various explanations and examine the impact of each type of explanation on team performance and trust. Also, discovering the most effective combinations of explanation would be worthwhile. Furthermore, it would be intriguing to gain insight into how the integration of individual differences in AI affects human trust, workload, and team performance. Lastly, rather than a computer-based simulator, building a tablet-based simulator with more painting tools may be an interesting experiment.

This study can be generalized in that it highlights the importance of AI learning sufficient individual differences during human-AI teaming tasks. This requires including individual differences in training data set, as well as enabling AI models to effectively learn individual differences. Additionally, the study showed that display transparency alone is not sufficient for understanding and building trust in AI systems.

REFERENCES

Allen, B., & Dreyer, K. (2018). The artificial intelligence ecosystem for the radiological sciences: ideas to clinical practice. *Journal of the American College of Radiology*, *15*(10), 1455-1457.

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., ... & Zhu, Z. (2016, June). Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning* (pp. 173-182). PMLR.

Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *new media & society*, *20*(3), 973-989.

Arlitsch, K., & Newell, B. (2017). Thriving in the age of accelerations: A brief look at the societal effects of artificial intelligence and the opportunities for libraries. *Journal of Library Administration, 57*(7), 789-798.

Bass, E. J., Baumgart, L. A., & Shepley, K. K. (2013). The effect of information analysis automation display content on human judgment performance in noisy environments. *Journal of cognitive engineering and decision making*, *7*(1), 49-65.

Bean, N. H., Rice, S. C., & Keller, M. D. (2011, September). The effect of gestalt psychology on the system-wide trust strategy in automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 55, No. 1, pp. 1417-1421). Sage CA: Los Angeles, CA: SAGE Publications.

Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zieba, K. (2016). End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*.

Bolstad, C. A., & Endsley, M. R. (1999, September). Shared mental models and shared displays: An empirical evaluation of team performance. In *proceedings of the human factors and ergonomics society annual meeting* (Vol. 43, No. 3, pp. 213-217). Sage CA: Los Angeles, CA: SAGE Publications.

Bolstad, C. A., Riley, J. M., Jones, D. G., & Endsley, M. R. (2002, September). Using goal directed task analysis with Army brigade officer teams. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 46, No. 3, pp. 472-476). Sage CA: Los Angeles, CA: SAGE Publications.

Boyce, M. W., Chen, J. Y., Selkowitz, A. R., & Lakhmani, S. G. (2015, March). Effects of agent transparency on operator trust. In *Proceedings of the Tenth Annual ACM/IEEE*

International Conference on Human-Robot Interaction Extended Abstracts (pp. 179-180).

Byrne, D. (1997). An overview (and underview) of research and theory within the attraction paradigm. *Journal of Social and Personal Relationships*, *14*(3), 417-431.

Cai, C. J., Jongejan, J., & Holbrook, J. (2019, March). The effects of example-based explanations in a machine learning interface. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 258-262).

Caldwell, S., Sweetser, P., O'Donnell, N., Knight, M. J., Aitchison, M., Gedeon, T., ... & Conroy, D. (2022). An Agile New Research Framework for Hybrid Human-AI Teaming: Trust, Transparency, and Transferability. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, *12*(3), 1-36.

Chan, W., Jaitly, N., Le, Q., & Vinyals, O. (2016, March). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 4960-4964). IEEE.

Chaudhri, V. K., Lane, H. C., Gunning, D., & Roschelle, J. (2013). Intelligent learning technologies Part 2: applications of artificial intelligence to contemporary and emerging educational challenges. *Ai Magazine*, *34*(4), 10-12.

Chen, J. Y., Lakhmani, S. G., Stowers, K., Selkowitz, A. R., Wright, J. L., & Barnes, M. (2018). Situation awareness-based agent transparency and human-autonomy teaming effectiveness. *Theoretical issues in ergonomics science*, *19*(3), 259-282.

Chen, T., Campbell, D., Gonzalez, F., & Coppin, G. (2014, December). The effect of autonomy transparency in human-robot interactions: a preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous UAV management. In *Australasian Conference on Robotics and Automation 2014* (pp. 1-10).

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., ... & Williams, M. D. (2021). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, *57*, 101994.

Endsley, M. R. (2001). A model of inter-and intrateam situational awareness: implications for design, training and measurement. *New trends in cooperative activities*, 46-68.

Endsley, M. R. (2017). From here to autonomy: lessons learned from human–automation research. *Human factors*, *59*(1), 5-27.

Endsley, M. R., Bolté, B., & Jones, D. G. (2003). *Designing for situation awareness: An approach to user-centered design*. CRC press.

Felzmann, H., Fosch-Villaronga, E., Lutz, C., & Tamò-Larrieux, A. (2020). Towards transparency by design for artificial intelligence. *Science and Engineering Ethics*, *26*(6), 3333-3361.

Gao, J., Lee, J. D., & Zhang, Y. (2006). A dynamic model of interaction between reliance on automation and cooperation in multi-operator multi-automation situations. *International Journal of Industrial Ergonomics*, *36*(5), 511-526.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Houssami, N., Lee, C. I., Buist, D. S., & Tao, D. (2017). Artificial intelligence for breast cancer screening: opportunity or hype?. *The Breast*, *36*, 31-33.

Khalid, H. M., Shiung, L. W., Nooralishahi, P., Rasool, Z., Helander, M. G., Kiong, L. C., & Ai-vyrn, C. (2016, September). Exploring psycho-physiological correlates to trust: Implications for human-robot-human interaction. In *Proceedings of the human factors and ergonomics society annual meeting* (Vol. 60, No. 1, pp. 697-701). Sage CA: Los Angeles, CA: SAGE Publications.

Khanna, S., Sattar, A., & Hansen, D. (2013). Artificial intelligence in health–the three big challenges. *The Australasian medical journal*, *6*(5), 315.

Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2021). Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *23*(6), 4909-4926.

Kunze, A., Summerskill, S. J., Marshall, R., & Filtness, A. J. (2019). Automation transparency: implications of uncertainty communication for human-automation interaction and interfaces. *Ergonomics*, *62*(3), 345-360.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human factors*, *46*(1), 50-80.

Lu, H., Li, Y., Chen, M., Kim, H., & Serikawa, S. (2018). Brain intelligence: go beyond artificial intelligence. *Mobile Networks and Applications*, *23,* 368-375.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of management review*, *20*(3), 709-734.

Mou, X. (2019). Artificial intelligence: investment trends and selected industry uses. *International Finance Corporation*, *8*.

Muhuri, P. K., Shukla, A. K., & Abraham, A. (2019). Industry 4.0: A bibliometric analysis and detailed overview. *Engineering applications of artificial intelligence*, *78*, 218-235.

National Academies of Sciences, Engineering, and Medicine. (2021). Human-AI Teaming: State-of-the-Art and Research Needs.

O’Neill, T., McNeese, N., Barron, A., & Schelble, B. (2022). Human–autonomy teaming: A review and analysis of the empirical literature. *Human factors*, *64*(5), 904-938.

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: An integrated meta-analysis. *Human factors*, *56*(3), 476-488.

Panganiban, A. R., Matthews, G., & Long, M. D. (2020). Transparency in autonomous teammates: Intention to support as teaming information. *Journal of Cognitive Engineering and Decision Making*, *14*(2), 174-190.

Parveen, R. (2018). Artificial intelligence in construction industry: Legal issues and regulatory challenges. *International Journal of Civil Engineering and Technology*, *9*(13), 957-962.

Rogers, S. K., Peterson, B., & Mendenhall, M. (2019). Autonomous Horizons: The Way Forward.

Schmitt, F., Roth, G., Barber, D., Chen, J., & Schulte, A. (2018). Experimental validation of pilot situation awareness enhancement through transparency design of a scalable mixed-initiative mission planner. In *Intelligent Human Systems Integration: Proceedings of the 1st International Conference on Intelligent Human Systems Integration (IHSI 2018): Integrating People and Intelligent Systems, January 7-9, 2018, Dubai, United Arab Emirates* (pp. 209-215). Springer International Publishing.

Selkowitz, A. R., Lakhmani, S. G., & Chen, J. Y. (2017). Using agent transparency to support situation awareness of the Autonomous Squad Member. *Cognitive Systems Research*, *46*, 13-25.

Seppänen, R., Blomqvist, K., & Sundqvist, S. (2007). Measuring inter-organizational trust—a critical review of the empirical research in 1990–2003. *Industrial marketing management*, *36*(2), 249-265.

Shively, R. J., Lachter, J., Brandt, S. L., Matessa, M., Battiste, V., & Johnson, W. W. (2018). Why human-autonomy teaming?. In *Advances in Neuroergonomics and Cognitive Engineering: Proceedings of the AHFE 2017 International Conference on Neuroergonomics and Cognitive Engineering, July 17–21, 2017, The Westin Bonaventure Hotel, Los Angeles, California, USA 8* (pp. 3-11). Springer International Publishing.

Wickens, C. D. (2009). The psychology of aviation surprise: An 8 year update regarding the noticing of black swans. In *2009 International Symposium on Aviation Psychology* (p. 1).

Wright, J. L., Chen, J. Y., & Lakhmani, S. G. (2019). Agent transparency and reliability in human–robot interaction: The influence on user confidence and perceived reliability. *IEEE Transactions on Human-Machine Systems*, *50*(3), 254-263.

Yang, J., Chen, Y., Huang, W., & Li, Y. (2017, September). Survey on artificial intelligence for additive manufacturing. In *2017 23rd international conference on automation and computing (ICAC)* (pp. 1-6). IEEE.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017, March). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 408-416).

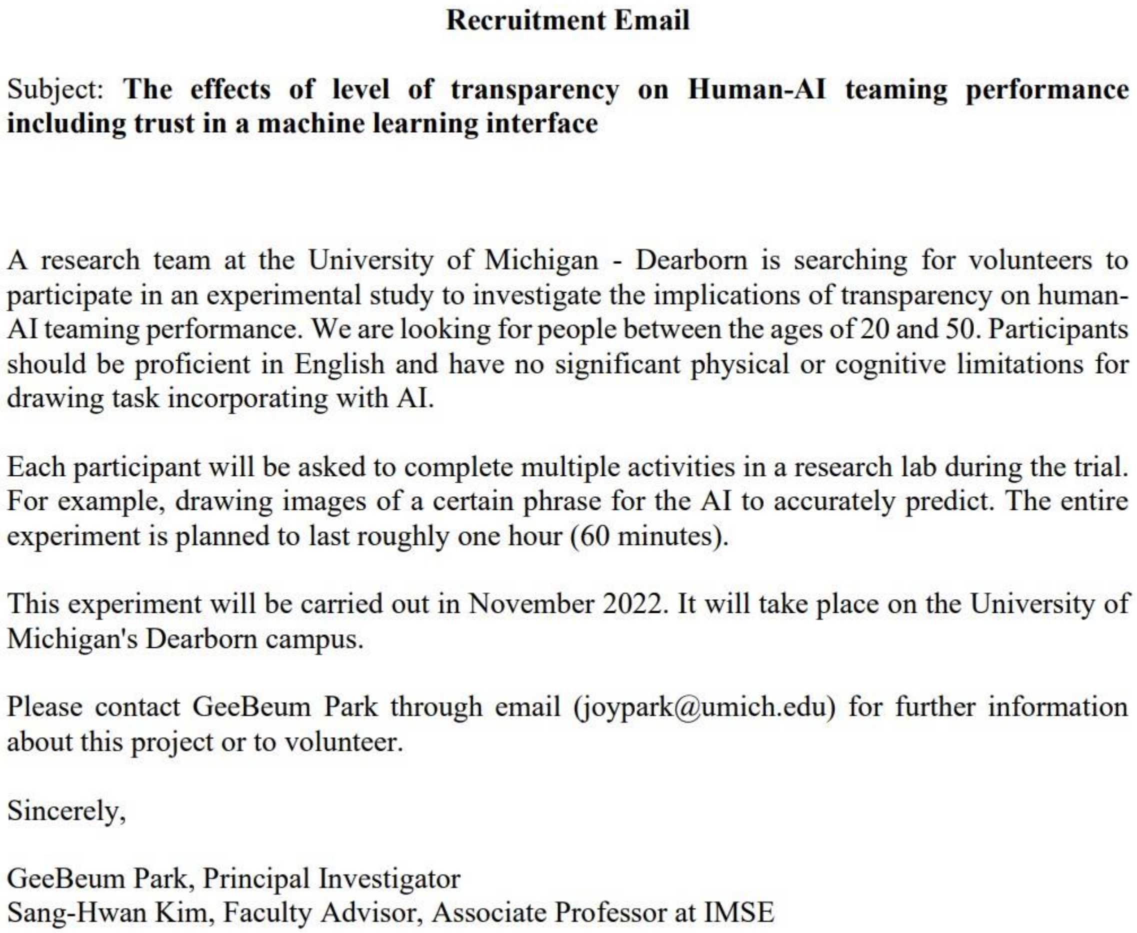
Yining, Y., & Zhiwei, X. (2018). doodleNet: A Family of Simple Convolutional Neural Networks for Image Classification Tasks. Retrieved from https://github.com/yining1023/doodleNet.

You, S., & Robert Jr, L. P. (2018, February). Human-robot similarity and willingness to work with a robotic co-worker. In *Proceedings of the 2018 ACM/IEEE International Conference on Human-Robot Interaction* (pp. 251-260).

Zhang, R., McNeese, N. J., Freeman, G., & Musick, G. (2021). " An ideal human" expectations of AI teammates in human-AI teaming. *Proceedings of the ACM on Human-Computer Interaction*, *4*(CSCW3), 1-25.

Zhong, R. Y., Xu, X., Klotz, E., & Newman, S. T. (2017). Intelligent manufacturing in the context of industry 4.0: a review. *Engineering*, *3*(5), 616-630.

APPENDIX A: Participants Recruitment Email



Recruitment Email

Subject: **The effects of level of transparency on Human-AI teaming performance including trust in a machine learning interface**

A research team at the University of Michigan - Dearborn is searching for volunteers to participate in an experimental study to investigate the implications of transparency on human-AI teaming performance. We are looking for people between the ages of 20 and 50. Participants should be proficient in English and have no significant physical or cognitive limitations for drawing task incorporating with AI.

Each participant will be asked to complete multiple activities in a research lab during the trial. For example, drawing images of a certain phrase for the AI to accurately predict. The entire experiment is planned to last roughly one hour (60 minutes).

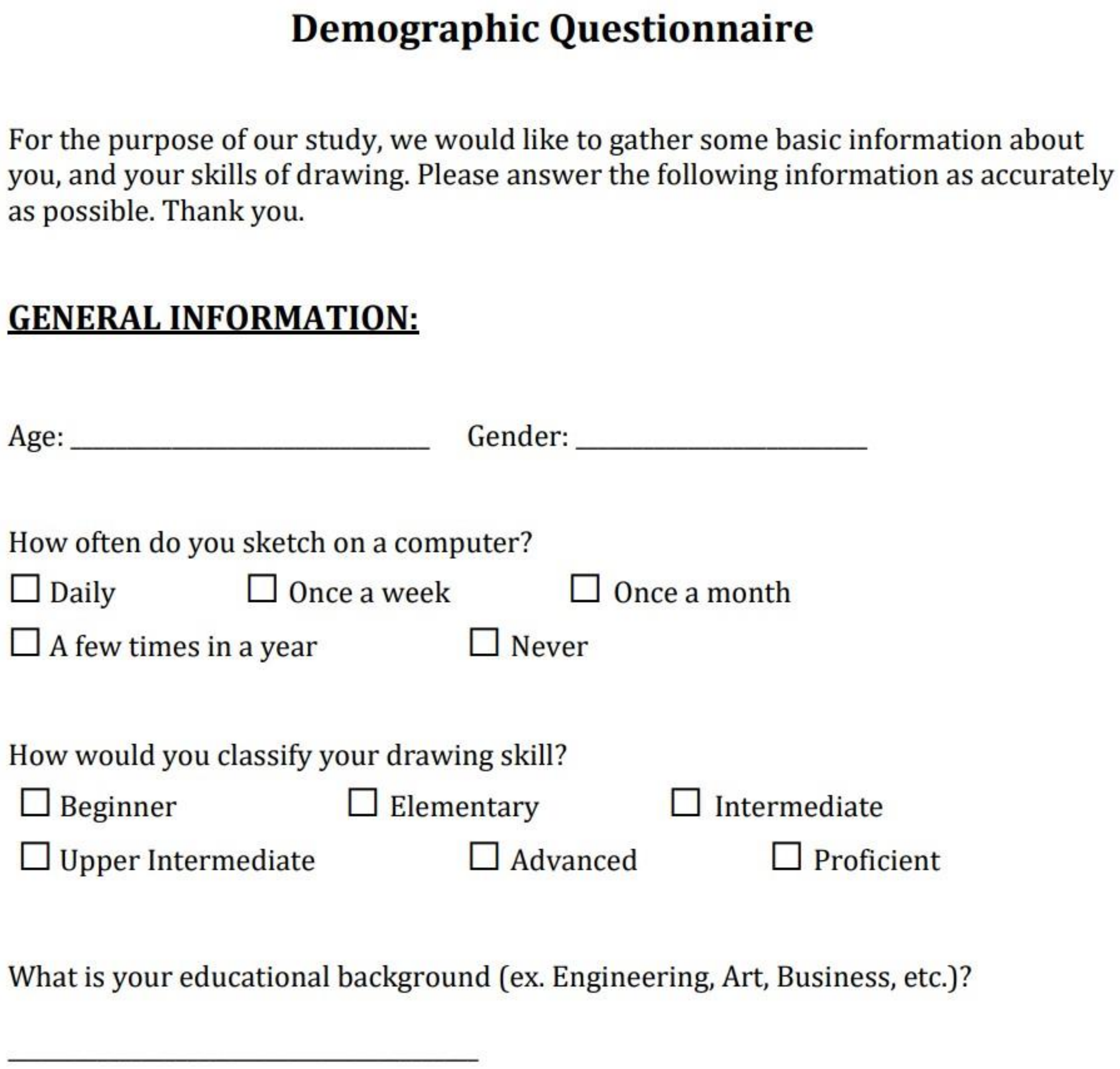
This experiment will be carried out in November 2022. It will take place on the University of Michigan's Dearborn campus.

Please contact GeeBeum Park through email (joypark@umich.edu) for further information about this project or to volunteer.

Sincerely,

GeeBeum Park, Principal Investigator
Sang-Hwan Kim, Faculty Advisor, Associate Professor at IMSE

APPENDIX B: Participants Demographic Questionnaire



Demographic Questionnaire

For the purpose of our study, we would like to gather some basic information about you, and your skills of drawing. Please answer the following information as accurately as possible. Thank you.

GENERAL INFORMATION:

Age: ______________________________ Gender: ________________________

How often do you sketch on a computer?

□ Daily □ Once a week □ Once a month

□ A few times in a year □ Never

How would you classify your drawing skill?

□ Beginner □ Elementary □ Intermediate

□ Upper Intermediate □ Advanced □ Proficient

What is your educational background (ex. Engineering, Art, Business, etc.)?

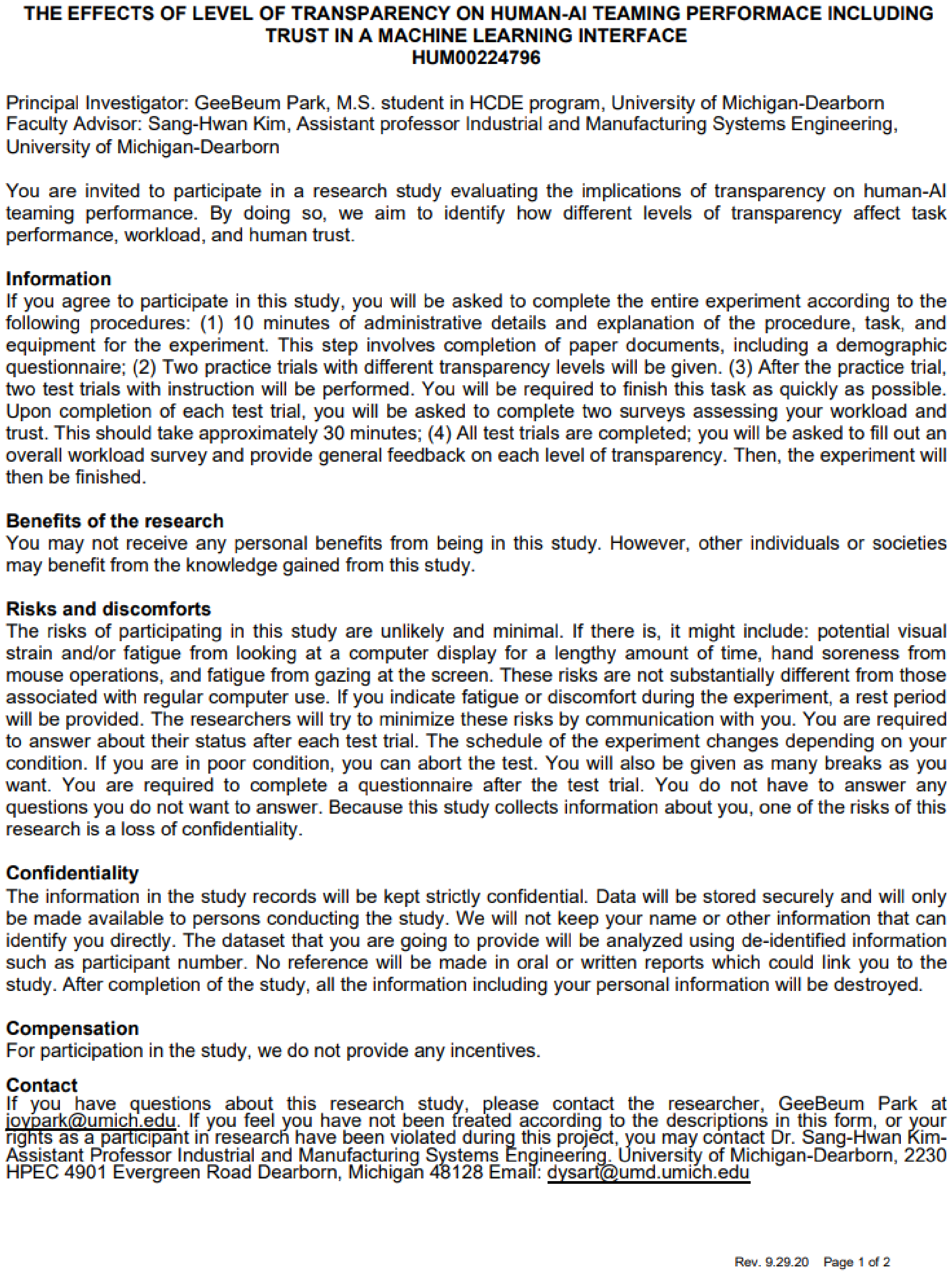
____

APPENDIX C: Participant Condition Assignments

	TRIAL #	
Subject #	**Experiment #1**	**Experiment #2**
1	A_5	B_1
2	B_5	A_1
3	A_1	B_5
4	B_1	A_5
5	A_5	B_1
6	B_5	A_1
7	A_1	B_5
8	B_1	A_5
9	A_5	B_1
10	B_5	A_1
11	A_1	B_5
12	B_1	A_5
13	A_5	B_1
14	B_5	A_1
15	A_1	B_5
16	B_1	A_5
17	A_5	B_1
18	B_5	A_1
19	A_1	B_5
20	B_1	A_5

A_1: System showing Top 1 Recognition with word set A
A_5: System showing Top 5 Recognition with word set A
B_1: System showing Top 1 Recognition with word set B
B_5: System showing Top 5 Recognition with word set B

APPENDIX D: Informed Consent Form



THE EFFECTS OF LEVEL OF TRANSPARENCY ON HUMAN-AI TEAMING PERFORMACE INCLUDING TRUST IN A MACHINE LEARNING INTERFACE
HUM00224796

Principal Investigator: GeeBeum Park, M.S. student in HCDE program, University of Michigan-Dearborn
Faculty Advisor: Sang-Hwan Kim, Assistant professor Industrial and Manufacturing Systems Engineering, University of Michigan-Dearborn

You are invited to participate in a research study evaluating the implications of transparency on human-AI teaming performance. By doing so, we aim to identify how different levels of transparency affect task performance, workload, and human trust.

Information

If you agree to participate in this study, you will be asked to complete the entire experiment according to the following procedures: (1) 10 minutes of administrative details and explanation of the procedure, task, and equipment for the experiment. This step involves completion of paper documents, including a demographic questionnaire; (2) Two practice trials with different transparency levels will be given. (3) After the practice trial, two test trials with instruction will be performed. You will be required to finish this task as quickly as possible. Upon completion of each test trial, you will be asked to complete two surveys assessing your workload and trust. This should take approximately 30 minutes; (4) All test trials are completed; you will be asked to fill out an overall workload survey and provide general feedback on each level of transparency. Then, the experiment will then be finished.

Benefits of the research

You may not receive any personal benefits from being in this study. However, other individuals or societies may benefit from the knowledge gained from this study.

Risks and discomforts

The risks of participating in this study are unlikely and minimal. If there is, it might include: potential visual strain and/or fatigue from looking at a computer display for a lengthy amount of time, hand soreness from mouse operations, and fatigue from gazing at the screen. These risks are not substantially different from those associated with regular computer use. If you indicate fatigue or discomfort during the experiment, a rest period will be provided. The researchers will try to minimize these risks by communication with you. You are required to answer about their status after each test trial. The schedule of the experiment changes depending on your condition. If you are in poor condition, you can abort the test. You will also be given as many breaks as you want. You are required to complete a questionnaire after the test trial. You do not have to answer any questions you do not want to answer. Because this study collects information about you, one of the risks of this research is a loss of confidentiality.

Confidentiality

The information in the study records will be kept strictly confidential. Data will be stored securely and will only be made available to persons conducting the study. We will not keep your name or other information that can identify you directly. The dataset that you are going to provide will be analyzed using de-identified information such as participant number. No reference will be made in oral or written reports which could link you to the study. After completion of the study, all the information including your personal information will be destroyed.

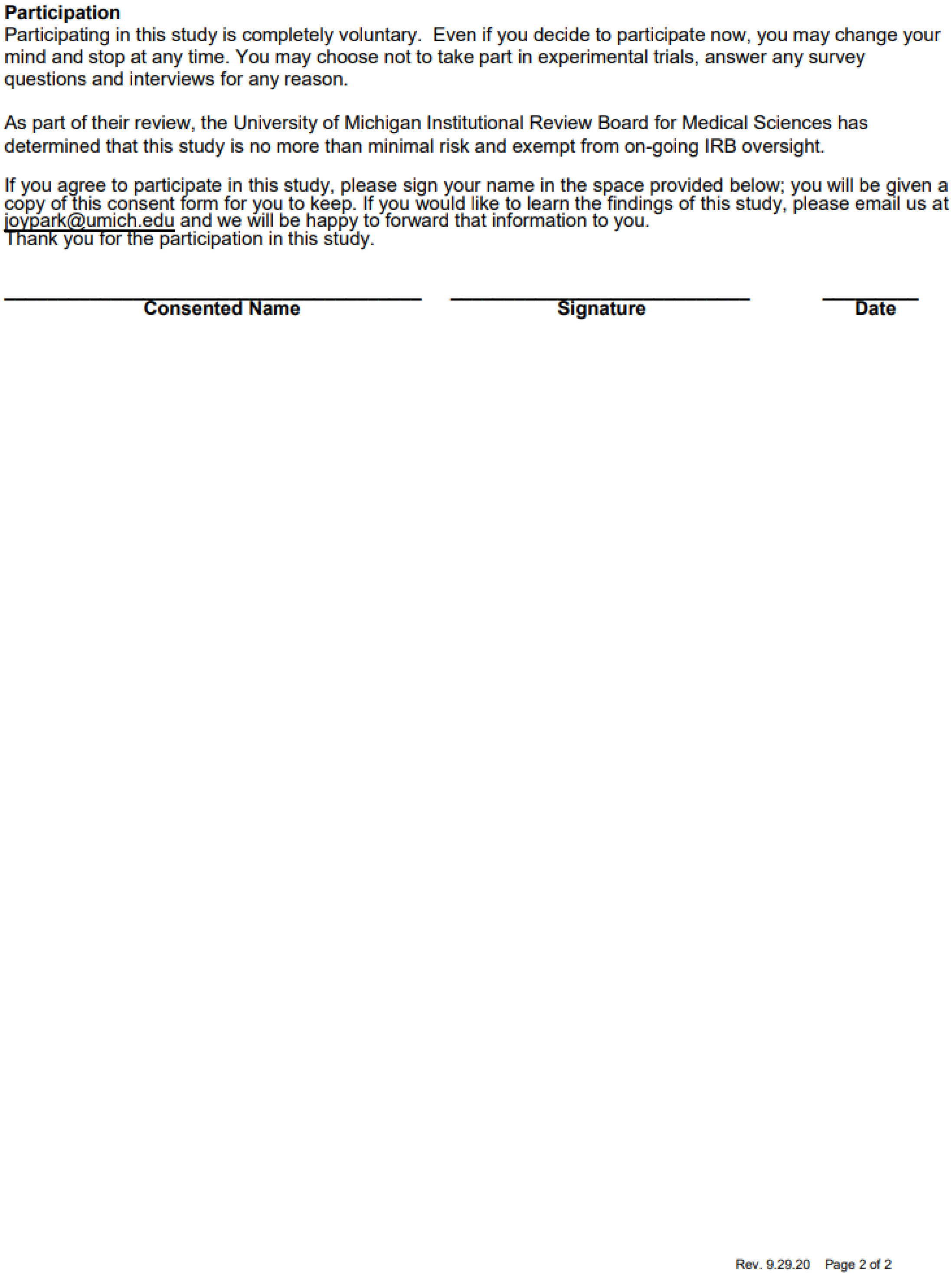
Compensation

For participation in the study, we do not provide any incentives.

Contact

If you have questions about this research study, please contact the researcher, GeeBeum Park at joypark@umich.edu. If you feel you have not been treated according to the descriptions in this form, or your rights as a participant in research have been violated during this project, you may contact Dr. Sang-Hwan Kim- Assistant Professor Industrial and Manufacturing Systems Engineering. University of Michigan-Dearborn, 2230 HPEC 4901 Evergreen Road Dearborn, Michigan 48128 Email: dysart@umd.umich.edu

Rev. 9.29.20 Page 1 of 2



Participation
Participating in this study is completely voluntary. Even if you decide to participate now, you may change your mind and stop at any time. You may choose not to take part in experimental trials, answer any survey questions and interviews for any reason.

As part of their review, the University of Michigan Institutional Review Board for Medical Sciences has determined that this study is no more than minimal risk and exempt from on-going IRB oversight.

If you agree to participate in this study, please sign your name in the space provided below; you will be given a copy of this consent form for you to keep. If you would like to learn the findings of this study, please email us at joypark@umich.edu and we will be happy to forward that information to you.
Thank you for the participation in this study.

______________________________ ______________________ _______
Consented Name **Signature** **Date**

APPENDIX E-1: NASA-TLX Rating Survey Form

Subjective Comparison of Demand Factors: NASA-TLX Survey

The effects of level of transparency on Human-AI teaming performance including trust
in a machine learning interface

Indicate the level of demand experienced during drawing task incorporation with AI for each of following factors by drawing a straight vertical line on the scale directly below.

Mental Demand How mentally demanding was the task?

Low High

Physical Demand How physically demanding was the task?

Low High

Temporal Demand How hurried or rushed was the pace of the task?

Low High

Performance How successful were you in accomplishing what you were asked to do?

Good Poor

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?

Low High

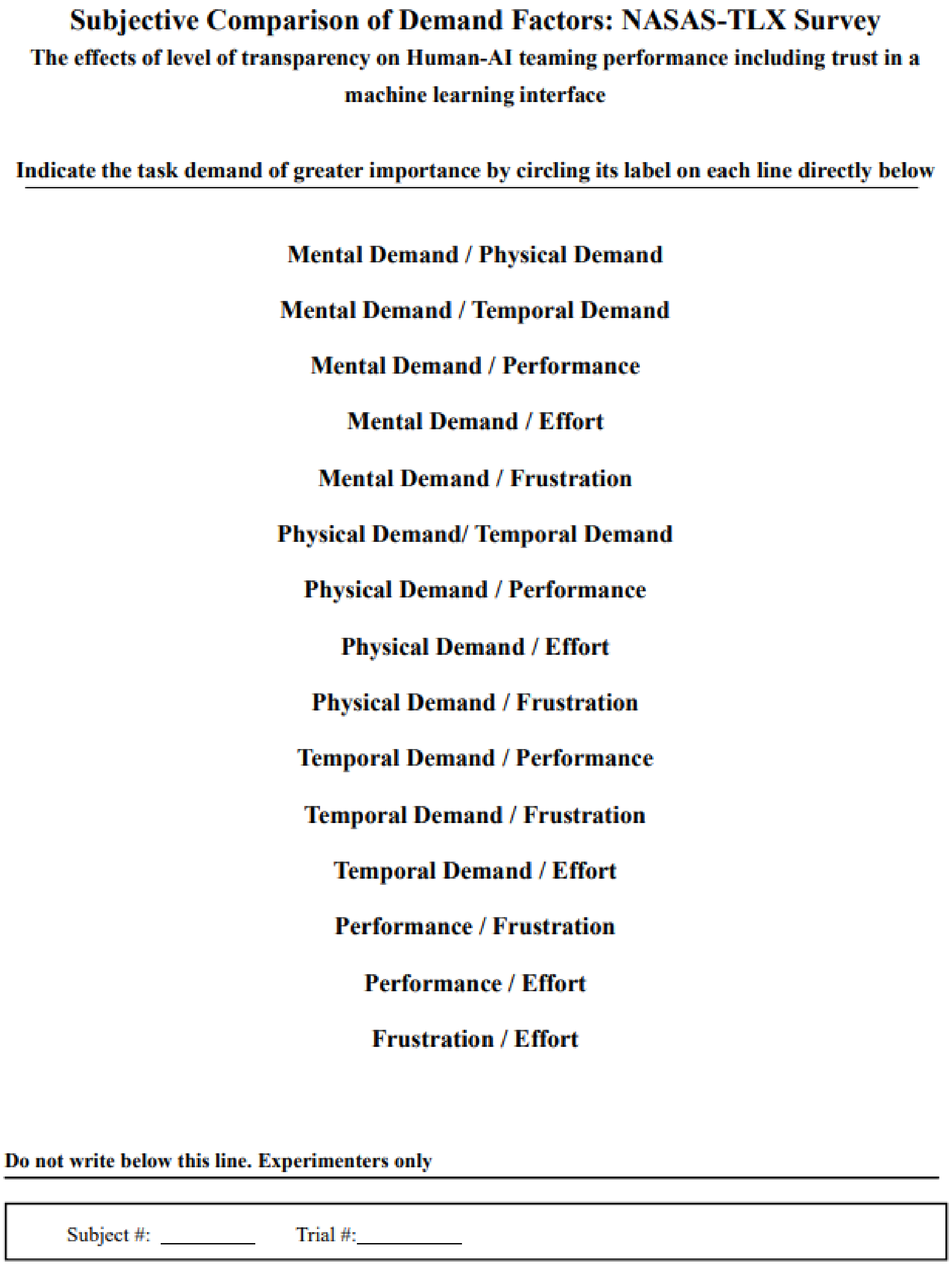
Effort How hard did you have to work to accomplish your level of performance?

Low High

Do not write below this line. Experimenters only

Subject #: _____ Trial #:

APPENDIX E-2: NASA-TLX Ranking Survey Form



Subjective Comparison of Demand Factors: NASAS-TLX Survey

The effects of level of transparency on Human-AI teaming performance including trust in a machine learning interface

Indicate the task demand of greater importance by circling its label on each line directly below

Mental Demand / Physical Demand

Mental Demand / Temporal Demand

Mental Demand / Performance

Mental Demand / Effort

Mental Demand / Frustration

Physical Demand/ Temporal Demand

Physical Demand / Performance

Physical Demand / Effort

Physical Demand / Frustration

Temporal Demand / Performance

Temporal Demand / Frustration

Temporal Demand / Effort

Performance / Frustration

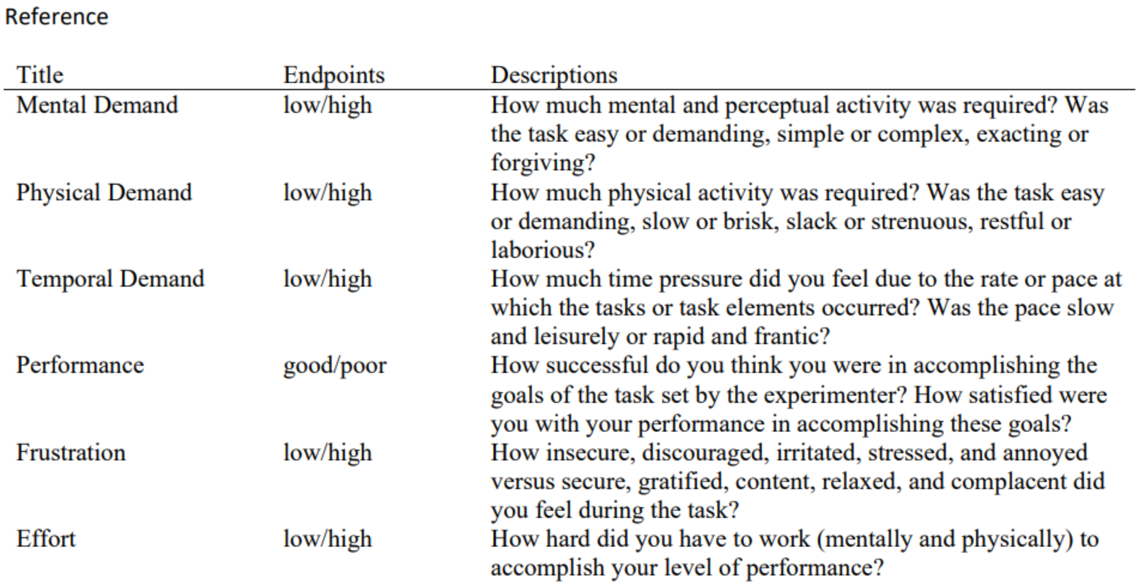
Performance / Effort

Frustration / Effort

Do not write below this line. Experimenters only

Subject #: _________ Trial #:__________

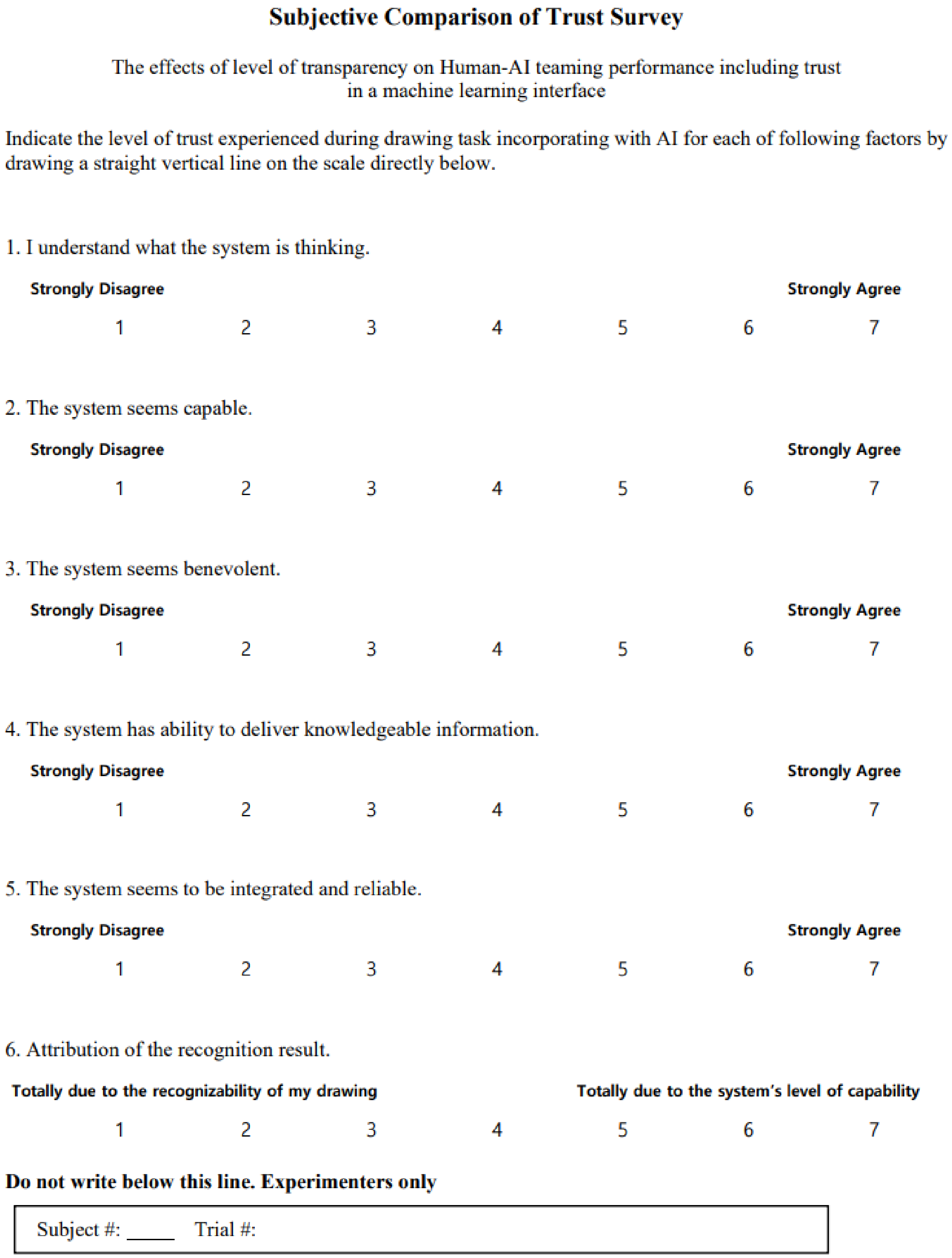
APPENDIX E-3: Definitions of Six Dimensions in NASA-TLX Survey



Reference

Title	Endpoints	Descriptions
Mental Demand	low/high	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex, exacting or forgiving?
Physical Demand	low/high	How much physical activity was required? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand	low/high	How much time pressure did you feel due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance	good/poor	How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?
Frustration	low/high	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?
Effort	low/high	How hard did you have to work (mentally and physically) to accomplish your level of performance?

APPENDIX F: Subjective Trust Survey Form



Subjective Comparison of Trust Survey

The effects of level of transparency on Human-AI teaming performance including trust
in a machine learning interface

Indicate the level of trust experienced during drawing task incorporating with AI for each of following factors by drawing a straight vertical line on the scale directly below.

1. I understand what the system is thinking.

Strongly Disagree **Strongly Agree**

1 2 3 4 5 6 7

2. The system seems capable.

Strongly Disagree **Strongly Agree**

1 2 3 4 5 6 7

3. The system seems benevolent.

Strongly Disagree **Strongly Agree**

1 2 3 4 5 6 7

4. The system has ability to deliver knowledgeable information.

Strongly Disagree **Strongly Agree**

1 2 3 4 5 6 7

5. The system seems to be integrated and reliable.

Strongly Disagree **Strongly Agree**

1 2 3 4 5 6 7

6. Attribution of the recognition result.

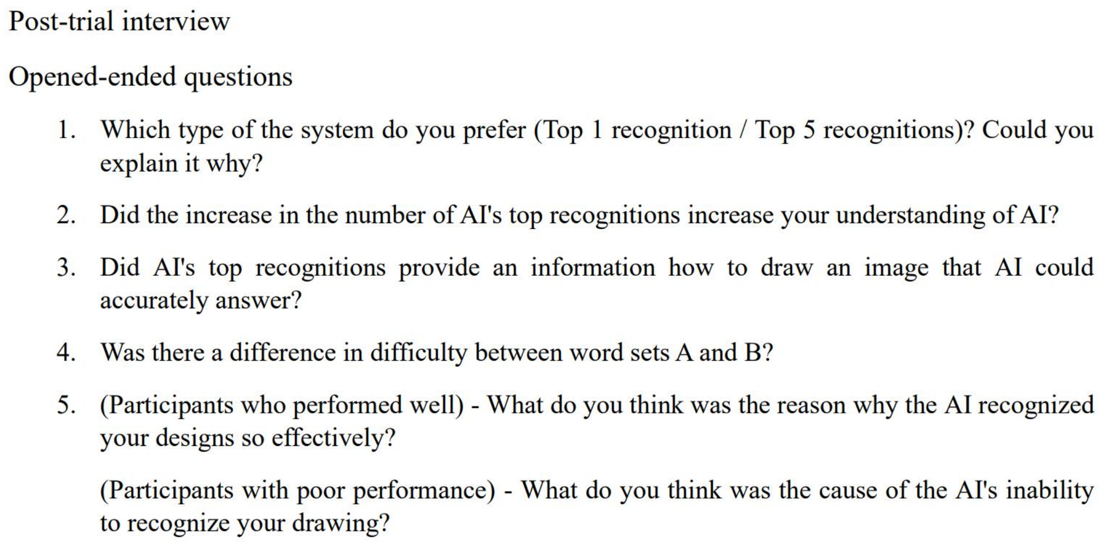
Totally due to the recognizability of my drawing **Totally due to the system's level of capability**

1 2 3 4 5 6 7

Do not write below this line. Experimenters only

Subject #: _____ Trial #:

APPENDIX G: Post-Trial Interview Questionnaire



Post-trial interview

Opened-ended questions

1. Which type of the system do you prefer (Top 1 recognition / Top 5 recognitions)? Could you explain it why?
2. Did the increase in the number of AI's top recognitions increase your understanding of AI?
3. Did AI's top recognitions provide an information how to draw an image that AI could accurately answer?
4. Was there a difference in difficulty between word sets A and B?
5. (Participants who performed well) - What do you think was the reason why the AI recognized your designs so effectively?

   (Participants with poor performance) - What do you think was the cause of the AI's inability to recognize your drawing?