An Investigation of Human Annotators' AI Teammate Selection and Compliance Behaviours

by

Jarvis Tse

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Mathematics
in
Computer Science

Waterloo, Ontario, Canada, 2024

Author's Declaration

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

Human-artificial intelligence (AI) collaborative annotation has gained increasing prominence as a result of its enormous potential to complement human and AI strengths as well as AI's recent development. However, it is not straightforward to form suitable human-AI teams and design human-AI interaction mechanisms for effective collaborative annotation. Through an exploratory study, this thesis investigated a diverse set of factors that may influence humans' AI teammate selection and compliance behaviours in a collaborative annotation context wherein AI agents serve as suggesters to humans. The study results indicate that multiple factors influenced which AI agents the participants chose to receive suggestions from, such as the AI agents' recent and overall accuracies as well as the participants' suggestion compliance records. We also discovered that the participants' AI compliance decisions were swayed by factors including whether the AI agents' suggestions aligned with the participants' top choices and whether such suggestions provided novel perspectives to the participants. Moreover, it was found that most of the participants constructed narratives to interpret the differences in various AI teammates' behaviours based on limited evidence. This thesis also contributes by presenting *MIA*, a versatile web platform for mixed-initiative annotation. Based on the weaknesses of *MIA*'s current designs, as informed by empirical results of the aforementioned exploratory study and another human-AI collaborative annotation study, as well as the goals to improve *MIA*'s scalability and adaptability, this thesis proposes design changes to *MIA*; these design changes also apply to other mixed-initiative annotation platforms.

Acknowledgements

I want to give my deepest thanks to my supervisor, Dr. Edith Law, for everything she has taught me as well as her generous encouragement, guidance, and patience throughout the process of completing this thesis. I am also grateful to all the collaborators and instructors I worked with during my master's program whose assistance made this thesis possible. In addition, many thanks to my thesis readers, Dr. Jian Zhao and Dr. Kate Larson, for their valuable feedback on this thesis. Finally, I want to thank my family for their unconditional support throughout my academic career.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

Artificial intelligence (AI) models that assist human decisions by providing potential answers (i.e., AI suggesters) have demonstrated their ability to enhance human productivity by improving work speed and accuracy for annotation tasks [47]. This collaborative synergy between humans and AI suggesters who act as humans' teammates in human-AI teams has been shown effective in diverse domains, including clinical decision-making [117], cellular labelling [205], image segmentation [3], in addition to numerous other applications [101]. This collaborative framework leverages the complementary strengths of human expertise and AI capabilities. Particularly, human experts contribute to their AI teammates' refinement and learning acceleration by validating and rectifying AI-generated output [57]. Simultaneously, novice individuals could benefit from observational learning, a well-established phenomenon [106, 132, 114, 228], by observing their AI teammates' task execution.

Despite the prevalence of human-AI collaboration for annotation tasks, little is known about how humans perceive the characteristics and actions of their AI teammates in these human-AI teams. For instance, although abundant research has been conducted on trust in the past decades [201], transferring findings from human-human trust to human-AI trust is complex due to distinct perceptions of AI agents [75]; therefore, these existing findings cannot be directly applied to the human-AI context without additional verification. Human annotators' perception of their AI teammates also causes numerous unique challenges that obstruct effective human-AI collaboration. Specifically, studies have revealed that humans may under- or over-rely their teammates for various reasons including being unable to accurately gauge their AI teammates' abilities, being misled by deceitful AI teammates, and inherently distrusting AI teammates compared to human teammates [157, 209, 16, 16, 151]. Moreover, in this evolving landscape, multiple AI models may coexist with each AI model possessing distinct attributes such as accuracy level, bias-variance tradeoffs, and response time [204, 4, 200, 5]. As human-AI annotation task configurations become increasingly diverse, individuals may face choices while simultaneously collaborating with multiple AI agents, which further complicates the pursuit of the understanding of human perceptions of AI agents and effective human-AI team decision-making. Besides, notably, human-AI collaborative annotation is a sub-field of mixed-initiative interaction, which is defined as "a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time" [8]; therefore, human-AI collaborative annotation can be considered as a specific type of mixed-initiative annotation, where both AI agent and human agent are present.

In short, human-AI collaborative annotation is versatile and has a diverse range of applications and the configuration wherein AI agents work as suggesters to humans has been shown effective in complementing human and AI strengths. However, research gaps exist in the understanding of factors that affect human annotators' perceptions of their AI teammates, as the findings from relevant research in the human-human context are not directly transferable, and the evolving landscape that multiple AI agents exist in the same human-AI team further complicates the inquiry to human annotators' perceptions and behaviours in human-AI teams. Therefore, this thesis aims to investigate human annotators' behaviours and perceptions of their AI teammates in the context of human-AI collaborative annotation wherein multiple AI teammate options are available simultaneously and the AI teammates' sole purpose is to provide suggestions to human annotators. Particularly, this thesis focuses on the factors that influence human annotators' decisions regarding which AI agents they choose to receive suggestions from (i.e., AI teammate selection) and whether to accept the AI-generated suggestions (i.e., AI teammate compliance). Our thesis statement is as follows:

> In human-AI annotation tasks, humans' perception of their AI teammates is influenced by a wide range of factors besides the accuracy of the AI's suggestions. Understanding these factors will help inform the designs of mixed-initiative annotation systems, particularly ones wherein AI agents act as suggesters to humans.

1.2 Contributions

This thesis offers the following contributions regarding understanding and facilitating human-AI decision-making in the context of human-AI collaborative annotation:

- An analysis of factors that influence humans' AI teammate selection and compliance when provided more than one AI teammate option through a human-subject think-aloud study.
- A flexible mixed-initiative annotation platform, *MIA*, which can effectively facilitate human-AI collaborative annotation.
- A new proposed *MIA* design based on the results of the two studies with various configurations and interface designs as well as the goals to improve *MIA*'s scalability and adaptability. These new designs may also apply to other mixed-initiative annotation platforms.

To elaborate, the influential factors discovered through the human-subject think-aloud study, wherein the participants talked out loud about their decision-making processes while performing annotation tasks and were interviewed, include the implied order of AI teammates, the historical performance records and predictive reliability of AI agents, instances of prior erroneous acceptance or rejection of AI-generated suggestions by humans, the congruence between AI-generated suggestions and the preferred choices of humans, the suggestions' provision of novel and insightful perspectives, along with others. Additionally, our findings underscore other human perceptions of their AI teammates, such as sense-making, wherein the participants construct narratives to clarify the functional mechanisms of AI agents and justify their preferences for one agent over another.

Another contribution of this thesis is the design and implementation of *MIA*, a versatile web application for humans to perform image annotation tasks alongside AI agents. By adopting a modular design, *MIA* can be easily modifiable to adapt to different annotation tasks and paired with AI agents with different characteristics and capabilities. In addition to being a flexible mixed-initiative annotation platform for data collection, *MIA* is also an ideal platform to conduct human-AI collaborative annotation research. By design, *MIA* allows researchers to easily select various experimental parameters, as well as to log user activities during annotation tasks and embed questionnaires into the study workflows.

Moreover, we propose an improved design of *MIA* in this thesis. By doing so, we aim to rectify existing weaknesses in *MIA*'s current designs, which were discovered through the aforementioned human-subject think-aloud study (i.e., the exploration study) as well as a study with a different design that was deployed to show *MIA*'s capability as a mixed-initiative annotation platform (i.e., the feasibility study).

1.3 Outline

This thesis is organized as follows:

- Chapter 2 outlines relevant existing work about factors that may impact human-AI collaboration as well as such factors' influences.
- Chapter 3 describes the goals and design of the exploration study (i.e., this thesis' main study), as well as *MIA*, the used AI agents, and the participants that facilitated the study.
- Chapter 4 presents the experimental results collected through the exploration study.
- Chapter 5 discusses a proposal of *MIA*'s future extensions, the limitations of this thesis, as well as potential future work.
- Chapter 6 concludes the thesis by recapping the objectives, results, and implications of this thesis.

Chapter 2

Related Work

2.1 The Psychology of Help-Seeking and Help-Giving Behaviours

A plethora of work has investigated factors that influence whether people seek help from and give help to others, although most of this work focuses on the human-human context. Helping behaviour (i.e., help-giving behaviour), as defined in [194], is “actions intended to assist another person with a problem or to relieve distress” and altruistic helping is defined as “helping behavior motivated by concern for the person in distress.” There is an ongoing debate about whether helping behaviour is ever purely altruistic, i.e., any helping behaviour may in fact be motivated by both one’s own concern and by one’s concern for others [194]. Moreover, facilitating and encouraging help-seeking behaviour (i.e., learners asking for assistance) in the interactive online learning context has been shown to be crucial for enabling effective learning [6]. Therefore, help-seeking could play a similar crucial role in improving human performance in the human-AI collaboration context. Below, we list discovered factors that may affect humans’ help-seeking and help-giving behaviours in both human-human and human-AI contexts by grouping these factors into three categories, namely human traits, human perceptions, and cost-benefit analysis.

2.1.1 Human Traits

According to existing literature, traits that people possess may have a significant influence on their decisions regarding whether and when to seek and provide help. These traits include humans' personalities, neurological characteristics, as well as demographic attributes.

Personalities

Existing research has shown that personality is closely related to people's help-seeking and help-giving behaviours. For example, people who are more agreeable and other-oriented empathetic have been shown to be more likely to provide help to others [167]. Also, conscientiousness has been linked to a higher possibility of help-seeking, whereas people who are more neurotic tend to avoid seeking help from others [154]. In addition, shyness may lead to less possibility of seeking help from people of different sex [46]. Past studies have also shown that people with different personality traits may interact with AI agents differently. For instance, Li et al. [123] have investigated how human personalities affect people's interaction with AI agents with different characteristics; according to their results, being more conscientious correlates to a higher possibility of trusting AI agents; also, open-mindedness correlates to trust in more cheerful AI agents. These discoveries indicate that both human personalities and AI characteristics may influence humans' help-seeking and help-giving behaviours towards AI agents, as trust is closely related to these two types of behaviours.

Neuroscience

Past neurology studies have found that people's prosocial behaviour is closely related to the activation of different brain regions. Particularly, the dorsal posterior cingulate cortex, ventromedial prefrontal cortex, dorsolateral prefrontal cortex, and midcingulate cortex were found to be consistently activated by prosocial behaviour [20]. Moreover, the likelihood that people engage in prosocial behaviour may be predicted using both affective and cognitive empathy, which activate different brain regions, such as the anterior midcingulate cortex and ventromedial prefrontal cortex, respectively [193].

Demographic Attributes

Previous studies have shown that people's demographic attributes, including religion, age, and gender, may impact their help-giving behaviours. For example, compared to more religious people, compassion may play a more important role in driving less religious people's help-giving behaviour [182]. Also, age may correlate to altruistic help-giving behaviour;

specifically, older people are more likely to put in effort for others altruistically compared to their younger counterparts [129]. In addition, gender may affect people's help-giving behaviour. In the online video game context, for example, the perceived genders of both the help provider and receiver may affect how likely help is provided [211]. Gender may also correlate to the types of help one provides for others (i.e., generally, men provide physical help more often and women provide mental help more often) [167].

2.1.2 Human Perceptions

Humans' help-seeking and help-giving behaviours may also be affected by their perceptions—including how they perceive the AI agents, themselves, among other things.

Human Perceptions of the AI Agents

Existing research shows that perceiving AI agents as possessing different personalities may vary how people behave in human-AI teams. For instance, the participants were more willing to confide in and listen to a serious, assertive AI agent compared to a cheerful agent in a previous study [123]; this fact hints that people may be more likely to seek help from more serious AI agents. This result also coincides with the finding from another study—chatbots should be concise, clear, and not use humour in conversations such that they can effectively assist humans [196]. Moreover, stated accuracy, stated confidence, and observed accuracy of AI agents may affect how much people are willing to comply with AI agents; specifically, the observed accuracy has the highest impact, followed by stated accuracy, which has a significant impact only when people do not have the opportunity to directly observe the AI agents' performance, and the impact of stated confidence is the lowest [223, 175]. Additionally, AI agents' provision of white-box explanations could enhance perceived AI assistance usefulness [38, 128]; however, this method may require additional time cost [38] and reinforce undesired human biases [128]. Lastly, embodiment has been identified as an effective method to improve the perceived trustworthiness of AI agents [92, 99].

Human Perceptions of Self

People's perceptions of themselves have been discovered to influence whether they decide to carry out help-seeking and help-giving behaviours. For instance, a driving force of help-giving behaviour is people's perceived moral obligation [213, 152]. Also, people may give help when doing so fits their self-identity [1]. Moreover, mood management—i.e., to make one feel good about oneself—has been shown to dictate whether a person decides to make a donation [48]. Besides, the presumed possession of the best solution [158] and high self-estimated competency [158, 212] have both been deemed to be causes of people avoiding help-seeking.

Other Types of Human Perceptions

Past studies have found people's perceptions other than of self and AI agents may also change their helping-seeking and help-giving behaviours. For instance, people may be more likely to seek help from peers who have equal status [116]. Another such perception is the helper's empathy towards the person being helped; for example, the amount a person donates to another person may positively correlate to empathy the donator feels toward the donatee [48]. Relevantly, prosociality has been shown to be trainable through the practice of perspective-taking (e.g., imagining the experience of others), which improves one's empathy towards others [183]. In addition, the perception of the existence of other potential help providers may decrease one's willingness to provide help due to diffusion of responsibility [167]. Also, when other potential help providers do not display the intention to give help, people may assume help-providing is unnecessary, this effect is called pluralistic ignorance [167]. Besides, the timing of help-seeking could be improved through prompting; some effective prompting methods include tutorial video [196], instructional prompts [187, 199], and direct messages [196].

2.1.3 Cost-Benefit Analysis

Psychological literature has established that cost-benefit analysis is an important determinant in humans' decision-making process [27, 28, 210, 192]. Some may even argue that all factors that affect help-seeking and help-giving behaviours can be interpreted as parts of cost-benefit analyses (i.e., whether it is beneficial to seek/provide help based on the estimated costs and benefits). Therefore, to limit the scope of the cost-benefit analysis factors, below we only discuss the factors that people interpret as costs and benefits of seeking and providing help in both the human-human and human-AI contexts.

Costs

Help-seeking and help-giving behaviours are constrained by different types of costs. Time and effort are costs that people consider for both help-seeking [83] and help-giving [167].

However, before deciding whether to seek help, people also consider the sacrifice of people's pride and self-esteem [41], which are not considered as costs of help-giving. On the other hand, people consider the expertise they contribute as cost only for help-giving [167]. Moreover, in the case where people decide to give help, generally, since there is limited time and resources, people must prioritize helping some causes over others [100].

Benefits

Benefits are a crucial motivation for both help-seeking and help-giving behaviours. People are generally prosocially apathetic and are not willing to exert significant effort to help others when there is no obvious benefit [130]. Notably, the benefits of help-giving can come in various forms: egoistic motivation is one of the most important benefit-related elements that drive people to help others [167]; people may also be motivated to give help when they believe others will help them back later—this type of help-giving behaviour is called reciprocal altruism [167, 211]; moreover, people may provide help due to community, instead of individual, interests [213]; furthermore, doing good by satisfying others' needs [147], feeling good for helping others [60], and looking generous to others [67] may all be considered as benefits of giving help; lastly, although some may argue people could provide help completely altruistically (i.e., purely due to others' benefits) [211], others believe the opposite—seemingly altruistic behaviour may be based on less obvious motives including empathy and avoiding feeling guilty [167]. Unlike help-giving, when people seek help there are usually clearly perceivable benefits since the aim of help-seeking is to accomplish help seekers' own goals (e.g., to perform better academically [43] and to gain financial benefits [70]). Therefore, it is crucial that the humans in human-AI teams are guided to work towards maximizing the performance of the human-AI teams, instead of their individual performance, since the improvement of individual performance does not necessarily translate to better human-AI team performance [17].

Despite the fact that there exists a wide range of studies on the psychology of help-seeking and help-giving behaviours, these studies did not investigate humans' decisions regarding AI teammate selection and compliance. Nonetheless, past research has provided a basis for this thesis as well as future studies by proffering a list of potential factors that should be investigated in the human-AI team context.

2.2 Study Design Inspirations

To design a study to investigate determining factors that affect humans' perceptions and decisions regarding their AI suggester teammates in the context of human-AI collaborative annotation, we explored various fields including mixed-initiative annotation, interface agents, and observational learning. In addition, we explored existing literature about factors that influence humans' perceptions of their AI teammates in the context of human-AI decision-making. Specifically, we delved into research on how trust has been investigated in the context of human-AI decision-making; also, we investigated the workflows that have been previously deployed to facilitate human-AI decision-making as well as these workflows' effects. Below, we list our discoveries from the aforementioned topics and discuss how we utilized the existing findings to inform the study design for the exploration study, which is the main study that is discussed in this thesis.

2.2.1 Fields Relevant to Human-AI Collaborative Annotation

Versatility is a feature shared by mixed-initiative annotation, interface agents, and observational learning; particularly, abundant studies in vastly different contexts have been conducted in these three fields. Below, we outline the findings of these studies, with an emphasis on the diversity of the existing research.

Mixed-Initiative Annotation

Mixed-initiative annotation, a sub-field of mixed-interaction interaction, which is defined as "a flexible interaction strategy in which each agent (human or computer) contributes what it is best suited at the most appropriate time" [8], constitutes a versatile framework that takes varying forms and can adapt to various types of tasks. Numerous scholarly inquiries have explored the dynamic environments of mixed-initiative annotation. In one notable study, an AI assistant proffers suggested labels and ambiguous information to a human annotator, who subsequently renders a final decision based on the provided guidance [184]. Additional investigations have demonstrated AI systems' capacity to assist human users in data labelling, either by extending users' single labelling actions to multiple data records [14] or by offering label recommendations, thereby allowing users to consider the most probable responses [47]. Furthermore, AI systems have been shown to harness human annotations to enhance their knowledge and capabilities, as exemplified by a study wherein an AI agent collaboratively annotates anomalies in time series data alongside human users, thereby leveraging shape-matching techniques derived from computer graphics to refine its knowledge [127].

Interface Agents

The rapid advancement of AI and computational technologies has led to a proliferation of user-facing technologies embedded with artificial intelligence. Among these, interface agents, which are defined as "software entities actively facilitating a user's interaction with an interactive interface" [125] hold prominence due to their adaptability. Specifically, previous research has established the versatile utility of interface agents in facilitating a diverse spectrum of tasks, encompassing activities such as augmenting internet search processes and facilitating tutoring sessions, among others [149, 81, 91, 90, 153, 109, 24, 111].

Observational Learning

Previous studies have proposed systems that allow task performers without task expertise (i.e., crowdworkers) to observe others' work and shown such systems could enable crowdworkers to improve their performance. For example, crowdworkers could self-correct if the work platform enables them to view the majority or average answers from other crowdworkers [106]. Similarly, crowdworkers may improve their accuracy by viewing other crowdworkers' answers and their corresponding comparative frequencies [132]; however, observing high-quality answers does not guarantee performance improvement for crowdworkers [51]. Observational learning between crowdworkers and experts may also improve crowdworkers' performance; for instance, crowdworkers could improve their accuracy by viewing experts' feedback [132], especially when explanations are included [30]. Moreover, crowdworkers could also improve their task performance through paired discussion [37, 79], or by reviewing others' work independently and in teams [228].

2.2.2 Previously Investigated Human-AI Decision-Making Topics

Trust and workflow are two topics that have been broadly explored in the context of human-AI decision-making. According to existing studies, both appropriate trust between human and AI teammates and suiting workflows for their interactions are crucial for achieving effective human-AI decision-making.

Trust in Human-AI Teams

Prior research has extensively examined trust within human-AI collaborations from diverse angles. It has been observed that improvements in AI performance and expertise complementarity do not necessarily correlate with enhanced human-AI team performance [17, 227, 85, 122]. This demonstrates the critical role of trust in shaping team dynamics. Efforts have been made to quantify trust in human-AI collaborations [212, 85]. For instance, it has been found that humans engage in cost-benefit evaluations during interactions with AI teammates, aligning with established psychological literature on decision-making [27, 28, 210, 192]. These evaluations are influenced by various factors including human goals, motivation, risk perception, effort perception, task complexity, and cognitive attributes [110, 160, 185]. However, transferring findings from human-human trust to human-AI trust is complex due to distinct perceptions of AI agents [75]. Past studies have proposed to incorporate shared mental models [13] and theory of mind [44] as means to enhance trust between humans and AI by enabling them to better understand and predict each other's behaviours. Some obstacles to these approaches include identifying effective measurements of success [13] and the difficulty of building AI models that can simultaneously perform a wide range of tasks that have little correlation [44]. Also, literature from the field of explainable AI has proposed a wide range of methods, such as through the use of data visualizations and natural language, to provide insights into how AI models make their decisions to humans [7]. Additionally, studies have explored factors influencing trust in AI agents, including human personality [82], human self-perceived task confidence [212], AI agent embodiment [171, 92], AI claimed capability [223, 16], capabilities [223, 175, 224], and response speed [29].

Workflows in Human-AI Teams

Prior investigations have explored the influence of diverse workflow configurations on human perceptions of their AI teammates. Within these examined workflows, human and AI collaborators may take divergent roles or share overlapping responsibilities. To name a few example scenarios, AI agents may be tasked to allocate tasks to human team members [77]; alternatively, they may engage in collaborative task execution alongside human counterparts by either contributing suggestions subject to human validation or supplying answers integrated with human responses through algorithmic processes [202, 163, 155, 226, 101]. However, despite these endeavours, only few studies have explored the factors influencing human AI teammate selection and compliance [101, 225]. According to these limited number of studies, variations in AI interaction design, such as prolonged wait times for AI suggestions, the introduction of additional user actions (e.g., clicking a button) to solicit AI input, and the requirement to register provisional answers prior to receiving AI suggestions, have been proposed as mechanisms that stimulate analytical thinking and potentially enable humans to select better AI agent teammates and comply with more accurate AI suggestions; however, overuse of these designs may also deter humans from cooperating with their AI teammates [29, 162, 62].

Drawing upon the insights from fields relevant to human-AI decision-making, we decided to focus on investigating human-interface agent collaborative annotation tasks for this thesis due to the prominence of mixed-initiative annotation and interface agents. In addition, we decided to incorporate observational learning in the exploration study's design by giving the participants opportunities to view AI suggestions.

Moreover, while trust in human-AI collaborations has been extensively studied, the nuanced relationship among trust, AI teammate selection, and compliance remains an under-explored area; hence, we designed the exploration study to fill this research gap. Lastly, for the exploration study, we devised a human-AI team workflow that possesses the strategic aim of fostering analytical thinking among participants while maintaining their cooperative engagement, based on the human-AI interaction mechanisms proposed by the existing human-AI team workflow literature.

Chapter 3

Methodology

We initiated our inquiry for this thesis by developing *MIA*, a web application, to facilitate mixed-initiative annotation. We intended *MIA* to be a versatile annotation environment where human(s) can effectively accomplish various types of annotation tasks by collaborating with one or multiple human/AI teammates. We then conducted two studies with *MIA*. Particularly, the first study (i.e., the feasibility study) was deployed to illustrate *MIA*'s capability as a mixed-initiative annotation platform. This chapter will focus on the second study (i.e., the exploration study), which was a think-aloud study conducted on *MIA* wherein the participants engaged in Greek character identification tasks and were afforded the choice between two AI agents for each group of tasks; in this study, during the participants' collaboration with an AI agent, the participants received suggestions from the chosen agent, with the options to accept or reject said suggestions. This study sought to provide comprehensive responses to the following two central research questions (RQs):

- **RQ1:** What factors exert a discernible influence on the participants' preference to select a specific AI agent to work with?
- **RQ2:** What factors significantly contribute to the participants' bias to accept the suggestions offered by the AI agent they have elected to work with?

3.1 Design of *MIA*

Below, we will first discuss the designs of *MIA*. We then outline the design of the exploration study as well as the design of the AI agents used for the exploration study. To conclude

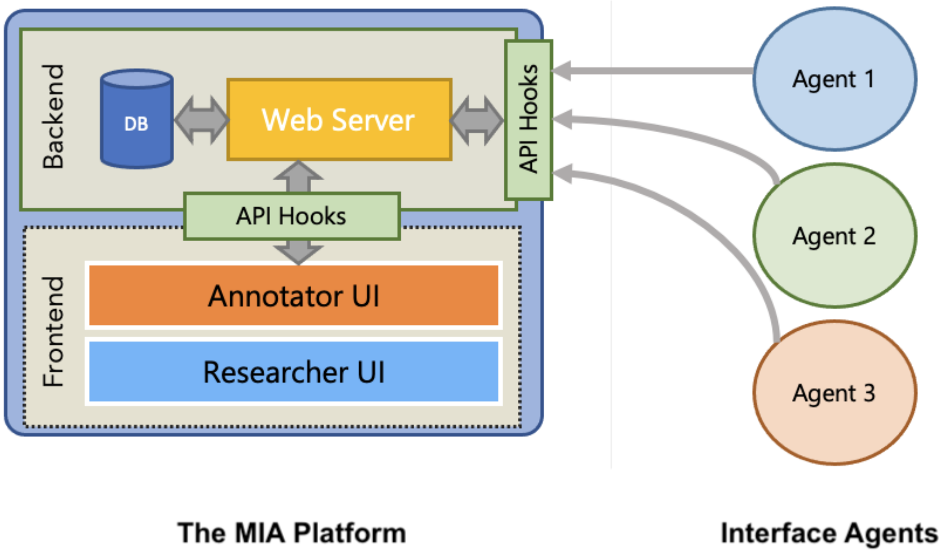


Figure 3.1: *MIA*'s System Architecture

this chapter, we examine the composition of the participants who participated in the exploration study.

3.1.1 *MIA*'s System Architecture

MIA is a web-based platform where practitioners and researchers can create and deploy annotation tasks to be completed by human annotators and/or machine learning models simultaneously. The system is implemented with a MeteorJS backend attached to a MongoDB database alongside a set of modular front-end components written in React. In addition to the standard functionality that often comes with web-based platforms (e.g., user account creation and management), *MIA* provides three key components to ease the burden of deploying human-AI team annotation tasks: (1) Application Programming Interface (API) hooks for activity observability, (2) a modifiable annotation interface for mixed-initiative annotation, and (3) an experimental task design module for designing and deploying mixed-initiative tasks. The system architecture is shown in Figure 3.1; note that in this figure the agent applications are shown to illustrate how they interface with the system; they are not part of *MIA*.

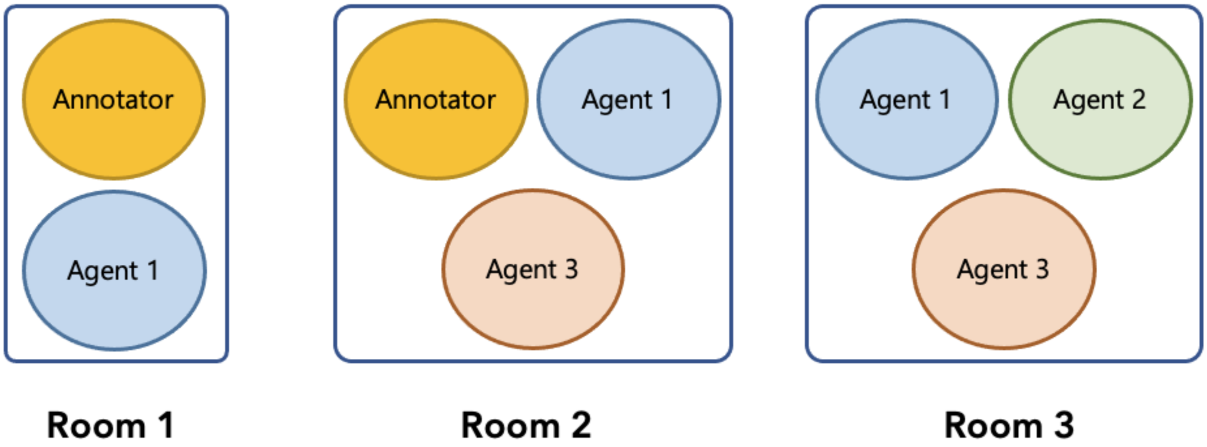


Figure 3.2: Example Mixed-Initiative Pairings of Annotators and Agents on *MIA*

Data Model: Annotation Rooms and API Hooks

Mixed-initiative interaction hinges on the abilities of human users and interface agents to adequately observe activities that take place within a given context. With this in mind, we implemented *MIA*'s architecture and data model around the notion of "*annotation rooms*," which we define as "annotation tasks in which annotation data, user activity, and user interface (UI) metadata are synchronized in real time among all present annotators." Our conceptualization of annotation rooms is motivated by the design of web-based messaging applications (e.g., chat rooms) that have similar requirements for observability and synchronization.

Our conceptualization of annotation rooms recognizes both human annotators and AI annotators as annotators that may be present. As shown in Figure 3.2, *MIA* allows researchers to engage a mix of human and AI annotators within a given annotation room. As each task creator's use case for the system may be unique, *MIA* does not systematically enforce the composition of human annotators and interface agents. For example, Figure 3.2's Room 3 illustration provides an example of how a room can be configured to allow several interface agents to interactively collaborate on a given annotation task.

MIA establishes a rich set of web-based API hooks to synchronize an annotation room's state among all human and AI annotators. All API hooks operate on the Datagram Delivery Protocol (DDP), a socket-to-socket delivery protocol standard that is specifically designed for real-time use cases. Collectively, *MIA*'s API hooks allow client applications (e.g., web browsers used by human annotators) and server applications (e.g., interface agents) not only to listen for activity relevant to a particular annotation room, but also contribute to the activity taking place within this same room. By default, the API hooks allow client and server applications to contribute arbitrary JSON data that is linked to an annotation room (i.e., by ID), which would then be broadcasted to the interfaces being used by other human or AI annotators in the same annotation room. The structural flexibility allows interfaces to send and query a variety of information relevant to the annotation room, such as annotation data, activity logs, and metadata about the task interface (e.g., width and height of the current image being annotated).

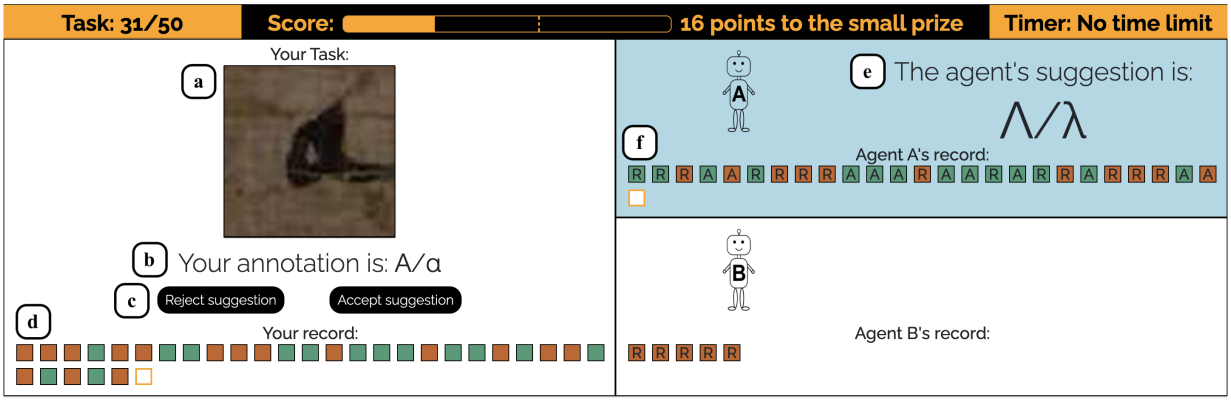


Figure 3.3: *MIA*'s Annotation Interface from the Exploration Study (labelled)

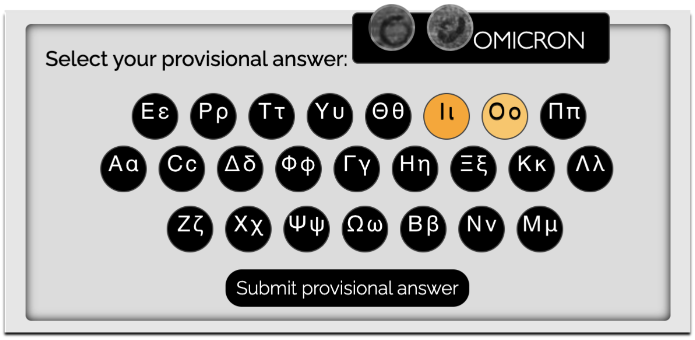


Figure 3.4: *MIA*'s Virtual Keyboard from the Exploration Study

3.1.2 Annotator UI: Completing Mixed-Initiative Annotation Tasks

To allow task designers to engage human annotators, *MIA* deploys its tasks for use with an annotation interface (Figure 3.3) that automatically facilitates synchronization via the system's API hooks. At load time, all relevant task information is provided to the annotation interface, and an event listener for the annotation room is then established to listen for room updates broadcasted by the server. The interface updates the UI to reflect any UI element changes (e.g., a new annotation) immediately upon their reception.

The annotation interface is modifiable to adapt different human-AI interaction mechanisms. For the exploration study, the annotation interface displays the following:

a. the annotated image;

b. the user's provisional answer;

c. the buttons for suggestion acceptance/rejection;

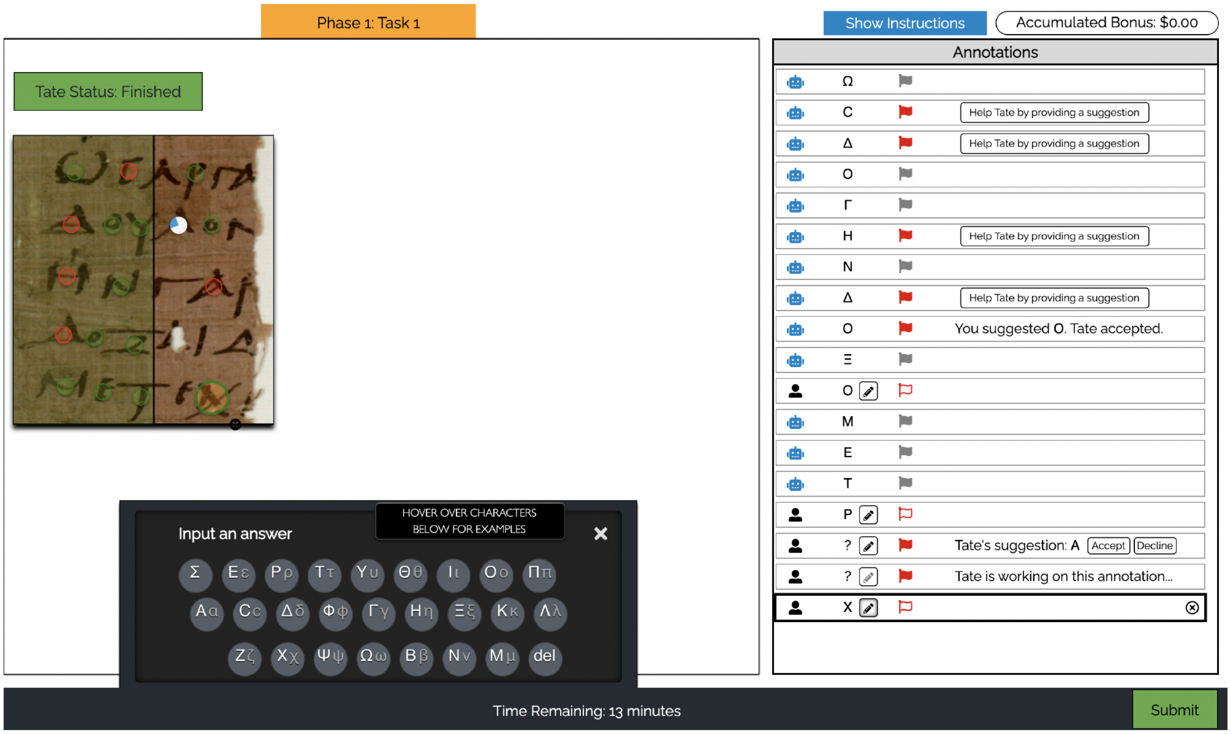
d. the user's task records (green and orange represent correct/incorrect final answers, respectively);

e. the chosen AI teammate's suggestion for the task;

f. each AI teammate's suggestion records (green and orange represent correct/incorrect suggestions, respectively; A's and R's represent whether the user accepted or rejected the corresponding suggestion, respectively).

This interface also contains a virtual keyboard (Figure 3.3), which takes user inputs for annotation labelling. Once a letter has been clicked the corresponding key becomes darker orange and the submit button becomes clickable. The user can also hover over each letter key on this virtual keyboard to view the corresponding examples, which will be displayed on the black panel on the top right part of the virtual keyboard, and the corresponding key will become light orange.

An alternative annotation interface is shown in Figure 3.5. For the exploration study, only one annotation was created for each annotated image and the AI agents were responsible for only providing suggestions to human annotators. However, *MIA*'s annotation interface can be modified to handle annotation tasks where multiple annotations are created for each image and each annotation is associated with 2-dimensional coordinates, in addition to a label. *MIA*'s annotation interface can also be modified to facilitate different human-AI assistance-seeking and assistance-giving mechanisms. For example, the interface


Phase 1: Task 1
Show Instructions
Accumulated Bonus: $0.00
Tate Status: Finished
Annotations
Ω
C
Help Tate by providing a suggestion
Δ
Help Tate by providing a suggestion
O
Γ
H
Help Tate by providing a suggestion
N
Δ
Help Tate by providing a suggestion
O
Ξ
O
M
E
T
P
?
?
X
Input an answer
HOVER OVER CHARACTERS BELOW FOR EXAMPLES
Σ Εε Ρρ Ττ Υυ Θθ Ιι Οο Ππ
Αα Cc Δδ Φφ Γγ Ηη Ξξ Κκ Λλ
Ζζ Χχ Ψψ Ωω Ββ Νν Μμ del
Time Remaining: 13 minutes
Submit

You suggested O. Tate accepted.

Tate's suggestion: A Accept Decline

Tate is working on this annotation...

Figure 3.5: An Alternative *MIA* Annotation Interface from the Feasibility Study

in Figure 3.5 enables humans to seek assistance with each annotation from an AI agent (i.e., *Tate*) by clicking on the annotation's corresponding flag icon located on the annotation panel on the right of the interface; also, this interface allows the AI agent to seek assistance from human users, who can respond to help requests by clicking the buttons on the annotation panel which are labelled "Help Tate by providing a suggestion." This interface design was adopted for *MIA*'s feasibility study, which we will further discuss in the Discussion section.

3.1.3 Task Creator UI: Deploying Mixed-Initiative Annotation Tasks

In order to allow creators, including practitioners and researchers, to leverage *MIA*'s utilities, *MIA* provides a set of user interfaces for designing and deploying mixed-initiative annotation tasks in three steps: (1) *Defining a Project*, (2) *Defining a Mixed-Initiative Annotation Task Design*, and (3) *Defining Experimental Instrumentation Use.*

Step 1: Defining a Project

"Projects" are *MIA*'s primary mechanism for creating and managing mixed-initiative annotation tasks. *MIA* allows task creators to create projects by uploading a CSV file that contains links to web-hosted media (i.e., images) that will be annotated by annotators. After each link has been parsed and added to *MIA*'s database as an individual data object, the task creator can conclude this step by providing a name for the project.

Step 2: Defining a Mixed-Initiative Annotation Task Design

Following a project's creation, *MIA* allows task creators to define mixed-initiative task designs and attach them to created projects for use. All task designs in *MIA* are orchestrated as task workflows composed of one or more "task phases" [98]. Users can specifically configure each phase's time limit, set of instructions, and task compositions (e.g., which images and in what order). Each of these attributes can be updated after task deployment.

Beyond these generic task design characteristics, *MIA* empowers researchers with substantial control over the mixed-initiative nature of its tasks. Task creators can specify an interface agent's participation in an annotation room by embedding the agent's `SECRET` API key as a parameter in their task design. When annotators subsequently load the annotation interface, the key will be forwarded to the client-side DDP library that forwards a notification to the AI agents' web server alongside any parameters relevant to the agents' behaviour.

Step 3: Defining Experimental Instrumentation Use

To support task creators, especially researchers, in using the platform as a tool to investigate human behaviour and perceptions in the context of mixed-initiative annotation, *MIA* provides several forms of instrumentation that administer research-related pop-up modals that annotators read, sign, or respond to during annotation tasks. These modals can include, for example, consent forms, feedback letters, and questionnaires (e.g., Google Forms). The system provides native support for engaging annotators before, during, and after specific work phases have ended as well as time-based intervals (e.g., every 10 minutes). Alongside these more explicit forms of instrumentation, *MIA* logs user interface activity (e.g., annotation events). Each of these logged activities is timestamped at the time of capture.

3.2 Study Design of the Exploration Design

This section discusses the design of the exploration study that was deployed on *MIA*. We start by describing the general design of the study, followed by the procedure for completing each individual annotation task during the study, the data collection and analysis methods, and finally, the composition of the study participants.

3.2.1 Overall Study Design

The study was conducted through in-person sessions, with the participants providing informed consent through an online form. Each participant undertook a series of 50 Greek letter annotation tasks, prefaced by verbal instructions. The participants were informed of a performance-based bonus incentive, in addition to a basic remuneration for completing all 50 tasks, to motivate their task engagement. Moreover, the participants were encouraged to engage in a think-aloud protocol, articulating their reasoning during task execution, to allow us to gain insights into their decision-making processes. Subsequent to task completion, the participants completed a demographic survey and engaged in a semi-structured interview aimed at clarifying the reasoning behind their decisions during the tasks. Finally, each participant received a feedback letter that clarified the study's objectives.

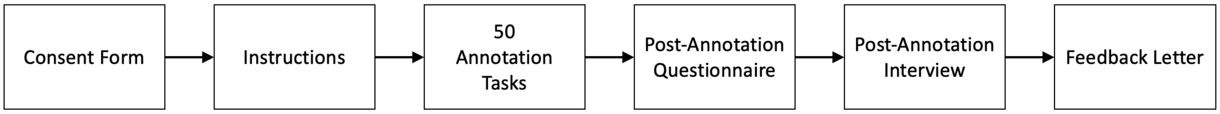


Figure 3.6: The Overall Workflow of the Exploration Study

3.2.2 Workflow for each Task of the Exploration Study

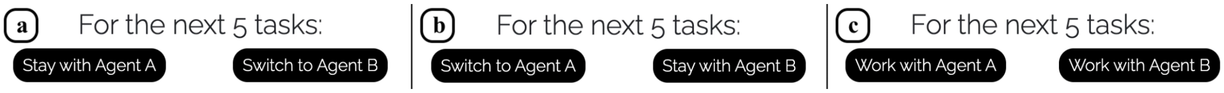
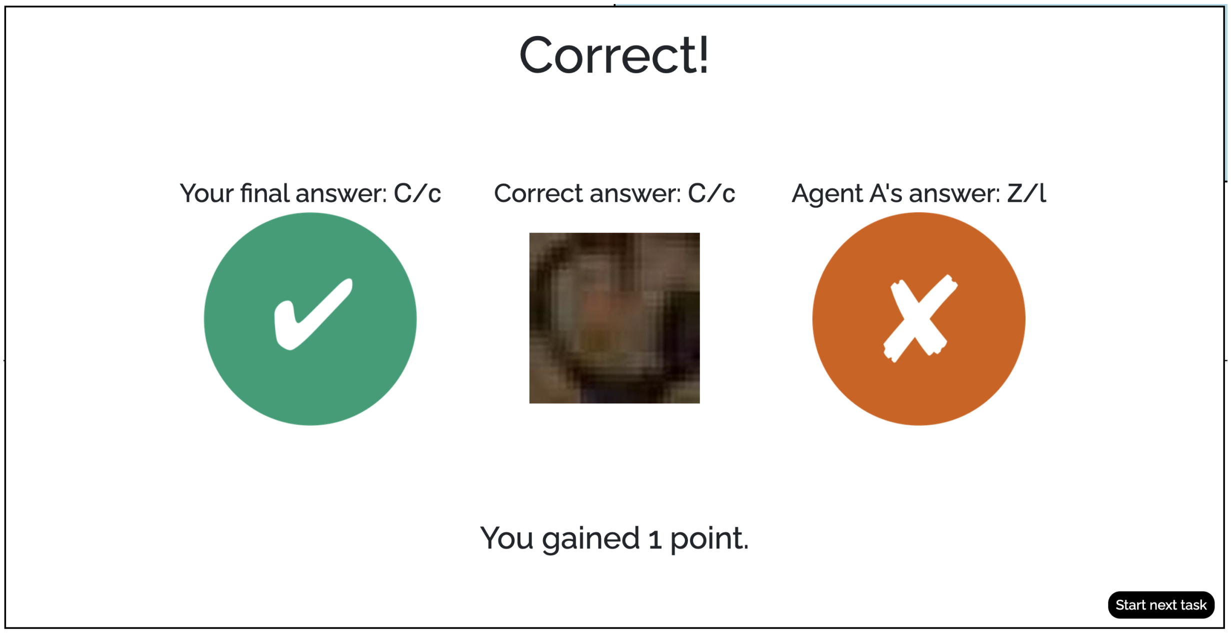


Figure 3.7: AI Agent Selection Button Pairs from the Exploration Study (labelled)

For each task, the participants were tasked with annotating a specific Greek letter present within an image excerpted from ancient Greek papyrus, as illustrated in Figure 3.3.a. These images, drawn from a pool of 75 selections within the AL-PUB_v2 dataset [197], were resized to conform to the dimensions of the interface, and each possessed an expert-validated ground truth annotation. Before the participants started annotating the first assigned image, they were requested to select their preferred AI agent collaborator out of two AI agents, designated as Agent A and Agent B, without knowing any performance-related information about these AI agents. Following the completion of each five tasks, the participants were provided with the opportunity to reevaluate their choice of AI agent for the subsequent set of five tasks. This choice was facilitated through the presentation of the corresponding pair of selection buttons, depicted in Figure 3.7; namely, components a, b, and c appeared for the participants who had just worked with Agent A, who had just worked with Agent B, and who just entered the first task, respectively. These selection buttons were displayed below the annotated image; this design permitted the participants to reference the performance records of both AI agents (Figure 3.3.f) as well as their own performance records (Figure 3.3.d), thus informing their decision-making process concerning agent continuation or alteration.


Correct!

Your final answer: C/c

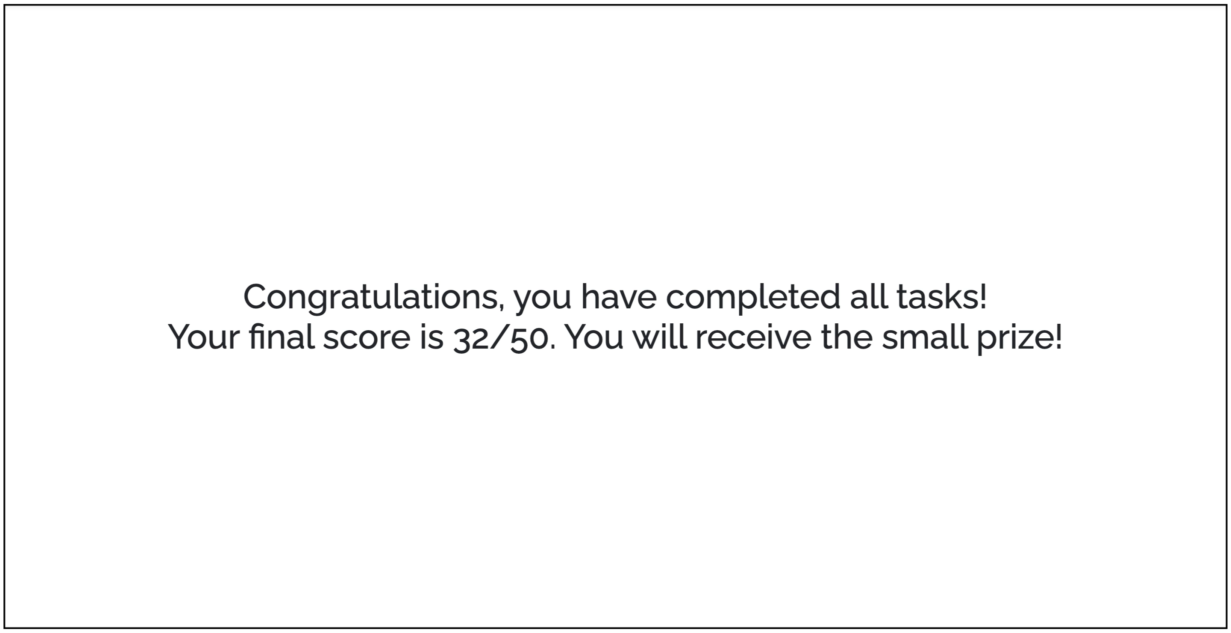
Correct answer: C/c

Agent A's answer: Z/l

You gained 1 point.

Figure 3.8: The Solution Pop-Up Window from the Exploration Study

Upon the selection of an AI agent, the agent would immediately start the annotation process for the current task; subsequently, the selected AI agent would start annotating for the corresponding task once each of the following four tasks was entered. Meanwhile, the participants specified their provisional answers for each task via a virtual keyboard



Congratulations, you have completed all tasks!
Your final score is 32/50. You will receive the small prize!

Figure 3.9: The End of All Tasks Screen from the Exploration Study

interface (Figure 3.4). This virtual keyboard appeared once a task was entered and was positioned at the bottom-center of the interface, below the interface components depicted in Figure 3.3, and it disappeared following the submission of a provisional answer. Following the provisional answer submission, the selected AI teammate's suggestion would be presented upon the completion of the AI annotation process for the corresponding task. The participants were allowed to either accept or reject the suggestion through the corresponding button in Figure 3.3.c. However, in instances where a participant's provisional answer coincided with the AI suggestion, the rejection option would remain unclickable, and the participant must accept the suggestion as the final answer. Once the submission of the final answer had been made for a task, a pop-up window (Figure 3.8) would appear. This window contained feedback including the user final answer's correctness, the agent suggestion's correctness, the annotated image, the change in user score, and a button to proceed to either the next task or the end-of-all-tasks screen (Figure 3.9), which displayed the final score of the participants. While immediate feedback provision may be unfeasible for certain real-world tasks, this limitation can be mitigated through asynchronous or estimation-based feedback mechanisms. Additionally, real-world tasks with ground truth data may allow the participants to evaluate their AI collaborators prior to commencing tasks without ground truth references. As such, the findings of this study hold relevance and applicability to practical human-AI annotation task scenarios.

3.2.3 Design of the Two AI Agents

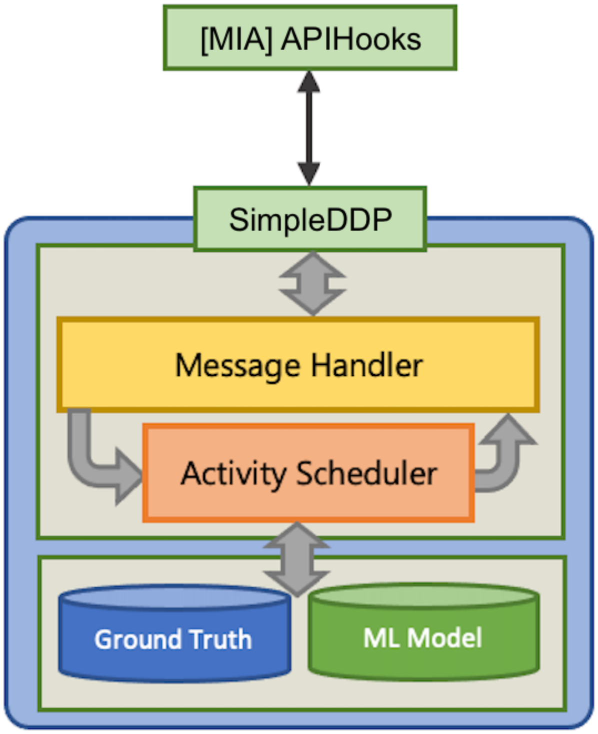


Figure 3.10: The AI Agents' Architecture

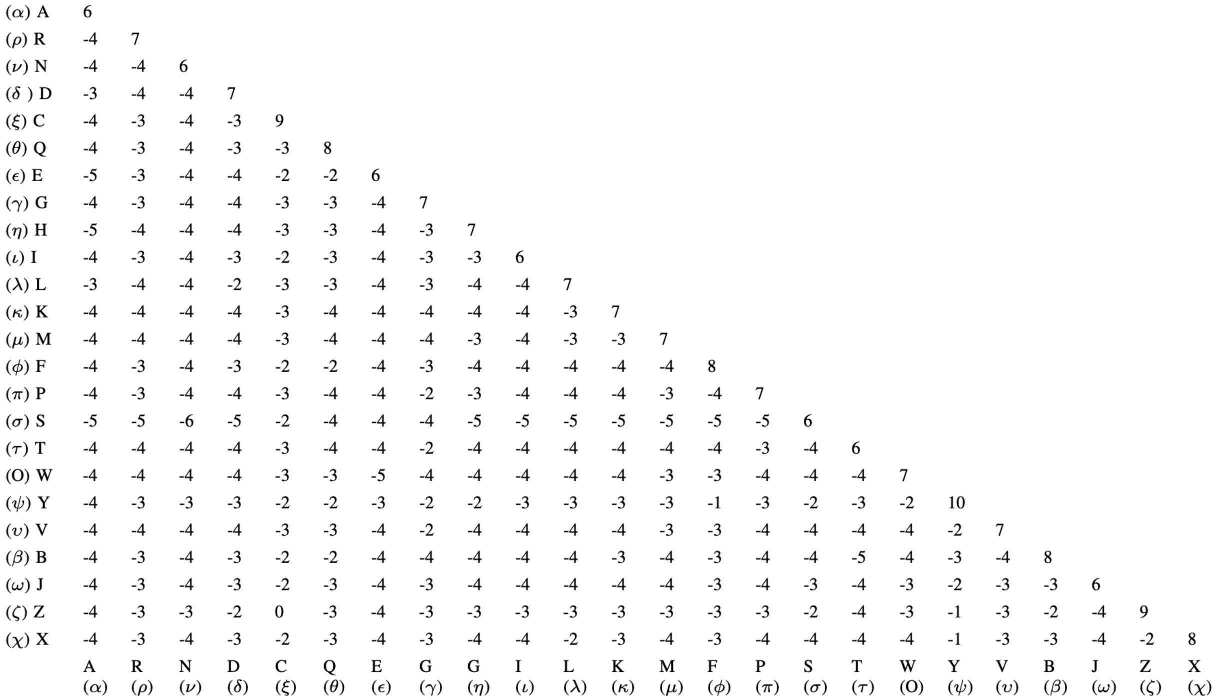
Both AI agents, which were implemented with Microsoft Bot Framework and Node.js, used the SimpleDDP JavaScript library to interface with *MIA*'s DDP-based API hooks. Architecturally, each AI agent can be sub-divided into two components:

1. **Message Handler:** A component that observes annotation room events sent through *MIA*'s API hooks and routes them to appropriate endpoints in the Activity Scheduler.

2. **Activity Scheduler:** A component that plans and schedules actions with messages received by the Message Handler. The component schedules two types of activities:

    (a) **Annotation Activity:** Activity that commands the AI agent to annotate its assigned image or image portions in a particular order, speed, etc.

(b) **Helping Activity:** Activity that commands the AI agent to provide assistance (i.e., feedback) with a collaborating annotator's annotation or seek assistance from a collaborating annotator with one of the AI agent's own annotations.



(α) A	6																							
(ρ) R	-4	7																						
(ν) N	-4	-4	6																					
(δ) D	-3	-4	-4	7																				
(ξ) C	-4	-3	-4	-3	9																			
(θ) Q	-4	-3	-4	-3	-3	8																		
(ϵ) E	-5	-3	-4	-4	-2	-2	6																	
(γ) G	-4	-3	-4	-4	-3	-3	-4	7																
(η) H	-5	-4	-4	-4	-3	-3	-4	-3	7															
(ι) I	-4	-3	-4	-3	-2	-3	-4	-3	-3	6														
(λ) L	-3	-4	-4	-2	-3	-3	-4	-3	-4	-4	7													
(κ) K	-4	-4	-4	-4	-3	-4	-4	-4	-4	-4	-3	7												
(μ) M	-4	-4	-4	-4	-3	-4	-4	-4	-3	-4	-3	-3	7											
(ϕ) F	-4	-3	-4	-3	-2	-2	-4	-3	-4	-4	-4	-4	-4	8										
(π) P	-4	-3	-4	-4	-3	-4	-4	-2	-3	-4	-4	-4	-3	-4	7									
(σ) S	-5	-5	-6	-5	-2	-4	-4	-4	-5	-5	-5	-5	-5	-5	-5	6								
(τ) T	-4	-4	-4	-4	-3	-4	-4	-2	-4	-4	-4	-4	-4	-4	-3	-4	6							
(O) W	-4	-4	-4	-4	-3	-3	-5	-4	-4	-4	-4	-4	-3	-3	-4	-4	-4	7						
(ψ) Y	-4	-3	-3	-3	-2	-2	-3	-2	-2	-3	-3	-3	-3	-1	-3	-2	-3	-2	10					
(υ) V	-4	-4	-4	-4	-3	-3	-4	-2	-4	-4	-4	-4	-3	-3	-4	-4	-4	-4	-2	7				
(β) B	-4	-3	-4	-3	-2	-2	-4	-4	-4	-4	-4	-3	-4	-3	-4	-4	-5	-4	-3	-4	8			
(ω) J	-4	-3	-4	-3	-2	-3	-4	-3	-4	-4	-4	-4	-4	-3	-4	-3	-4	-3	-2	-3	-3	6		
(ζ) Z	-4	-3	-3	-2	0	-3	-4	-3	-3	-3	-3	-3	-3	-3	-3	-2	-4	-3	-1	-3	-2	-4	9	
(χ) X	-4	-3	-4	-3	-2	-3	-4	-3	-4	-4	-2	-3	-4	-3	-4	-4	-4	-4	-1	-3	-3	-4	-2	8
	A (α)	R (ρ)	N (ν)	D (δ)	C (ξ)	Q (θ)	E (ϵ)	G (γ)	G (η)	I (ι)	L (λ)	K (κ)	M (μ)	F (ϕ)	P (π)	S (σ)	T (τ)	W (O)	Y (ψ)	V (υ)	B (β)	J (ω)	Z (ζ)	X (χ)

Figure 3.11: Greek Letter Oriented Substitution Matrix Adapted from [214]

For the exploration study, the two AI agents differed only in names and positions on the interface. Both AI agents had the sole responsibility of providing suggestions whenever they were selected by a human teammate to do so. Both agents required 3 seconds to complete each annotation (i.e., to provide a suggestion). The duration of 3 seconds was selected based on the pilot study results, such that the agents would act as if they needed time to think but would rarely require users to wait.

The performance of the two AI agents was controlled and randomized using identical protocols to ensure each participant would encounter diverse correctness/incorrectness sequences over the 10 sets of 5 tasks that they would work on. Particularly, both AI agents' behaviour was randomized using the same set of rules; hence, choosing which AI agent to work with had no effect on what suggestions the participants would receive. The correctness of each suggestion was determined before the participant entered the first task:

10 permutations were randomly picked from the potential 32 permutations of length 5 sequences with each term being either correct or incorrect; the 10 picked permutations were then concatenated into a sequence of length 50 in random order. This configuration allowed a diverse set of situations to occur by ensuring the AI agents' performance (i.e., correctness) varied over the course of the 50 tasks. Additionally, the letters the AI agents suggested were based on the ground truths when the AI agents made incorrect suggestions. Specifically, each letter had a corresponding similar-looking letter, which the AI agents would use as the incorrect suggestions; these corresponding letters were selected based on the Greek Letter Oriented Substitution Matrix [214] (Figure 3.11), a confusion matrix that quantifies the similarity level between each pair of Greek letters, wherein a greater non-positive number indicates stronger similarity.

3.2.4 Data Collection and Analysis

Each participant was accompanied by a study conductor throughout the annotation tasks, with the participants actively encouraged to vocalize their decision-making processes and engage in think-aloud protocols. The study conductor had the discretion to temporarily interrupt the participants during annotation tasks to solicit detailed explanations of their decisions. In the subsequent semi-structured interview conducted at the study session's culmination, the participants were probed regarding the determinants influencing their choices pertaining to AI teammate selection and compliance. Subsequently, the participants' verbal explanations and responses were subjected to thematic categorization, followed by summarization. Furthermore, the annotation interface logged participant and AI agent actions during the annotation tasks, as well as both the participants' and AI agents' correctness records. This included the annotations provided by the participants, whether they were correct, and the acceptance status of each AI suggestion. Quantitative analysis was applied to assess participant annotation behaviours and responses to the post-annotation questionnaire, which investigated the participants' demographic information. Given the repeated measurements acquired from the participants, we endeavoured to employ mixed-effect models whenever feasible, thereby enhancing statistical rigour and accounting for interdependencies within the dataset.

3.2.5 Participants

A total of 20 participants, whom we refer to as P1-20, were recruited via physical posters and through the research participant pool from the University of Waterloo. The participants received compensation in the amount of 15.00 Canadian dollars, complemented by

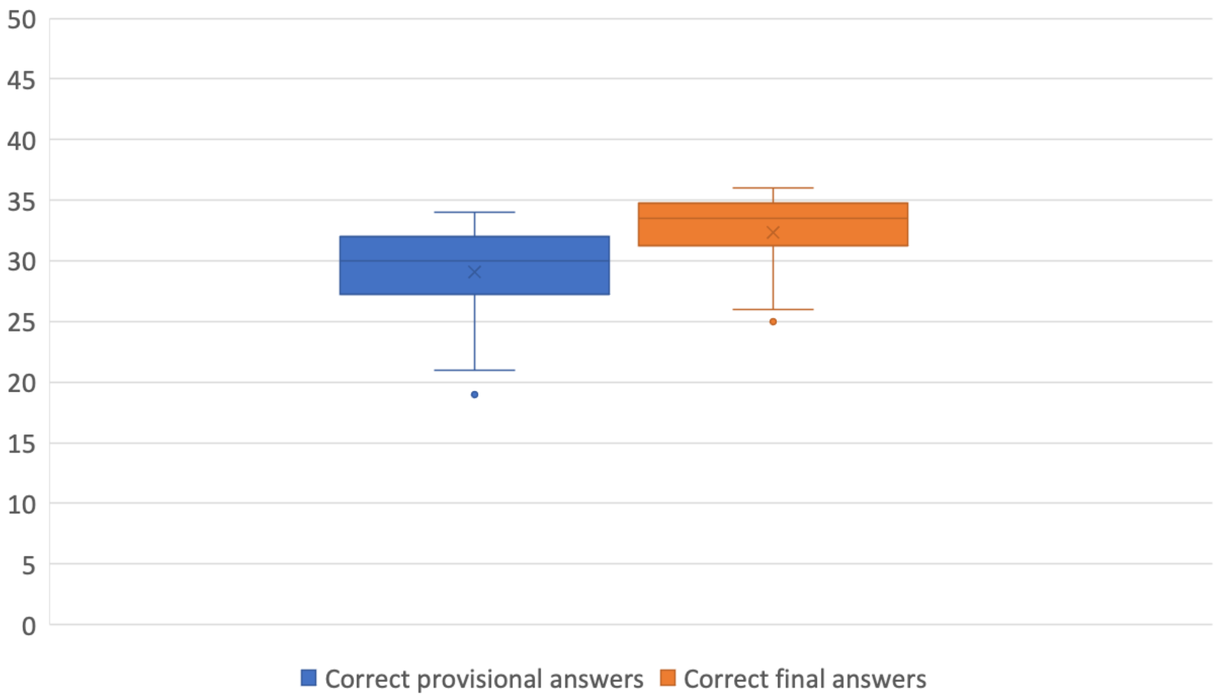


Figure 3.12: Correct Final and Provisional Answer Counts of the Exploration Study Participant

the prospect of two performance-based incentives, which were items priced 1.00 and 14.99 Canadian dollars, respectively. Of the 20 participants, 12 identified as female, 6 as male, and 2 as non-binary. Age demographics indicated that 12 participants were in the 18-24 age bracket, 7 fell within the 25-34 age range, and 1 participant was aged 35-44. The educational backgrounds of the participants were diverse, with 9 participants without a bachelor's degree, 6 holding a bachelor's degree as their highest level of education, and the remaining 5 participants having attained a master's degree. Proficiency in Egyptology, Assyriology, or Ancient Greek also exhibited a broad spectrum: 6 participants possessed no formal education in these fields, 6 had received K-12 education with related content, 6 had completed undergraduate coursework in the respective domains, and 2 participants, although possessed no relevant formal education, pursued these areas as a hobby. Performance metrics revealed that the participants achieved an average of 32.30 (SD = 3.18) correctly completed tasks out of 50 total tasks, with a minimum of 25 tasks. On average, the participants modified their provisional answers to align with AI teammate suggestions

ID	Overall Education	Relevant Education	Age	Gender
P1	High school diploma or GED	K-12	18 - 24	Female
P2	High school diploma or GED	None	18 - 24	Female
P3	Some college, but no degree	Undergraduate	18 - 24	Female
P4	High school diploma or GED	K-12	18 - 24	Male
P5	Bachelor's Degree	Undergraduate	18 - 24	Female
P6	Master's Degree	None	25 - 34	Male
P7	Bachelor's Degree	Undergraduate	25 - 34	Male
P8	High school diploma or GED	K-12	18 - 24	Female
P9	High school diploma or GED	Undergraduate	18 - 24	Female
P10	High school diploma or GED	None	18 - 24	Female
P11	Master's Degree	Undergraduate	25 - 34	Male
P12	High school diploma or GED	K-12	18 - 24	Non-Binary
P13	High school diploma or GED	K-12	18 - 24	Female
P14	Bachelor's Degree	K-12	18 - 24	Non-Binary
P15	Bachelor's Degree	None	18 - 24	Female
P16	Bachelor's Degree	Undergraduate	25 - 34	Female
P17	Master's Degree	None	25 - 34	Female
P18	Bachelor's Degree	Hobbyist	25 - 34	Male
P19	Master's Degree	None	25 - 34	Female
P20	Master's Degree	Hobbyist	35 - 44	Male

Table 3.1: Exploration Study Participant Demographic Information, Overall Education Level, and Education Level in Relevant Fields Including Egyptology, Assyriology, and Ancient Greek

9.80 times (SD = 3.55), with a maximum of 16 revisions. As shown in Figure 3.12, having the option to follow AI suggestions allowed the participants to improve their performance.

Chapter 4

Exploration Study Results

This chapter discusses the results collected from the exploration study. Particularly, below, we will first go over the relevant results from the exploration study that answers RQ1 and RQ2, respectively; then, we will outline other collected results from this study, including how the participants reacted to various human-AI performance scenarios and other noteworthy miscellaneous quotations.

4.1 Main Factors that Affected AI Teammate Selection (RQ1)

In our pursuit of validating the statistical significance of the discovered factors that affected the participants' AI agent selection, we deployed a generalized linear mixed-effect model (GLMM). However, our analytical efforts encountered singularity issues due to the small sample size and high consensus among the participants. While we were unable to unveil statistically significant outcomes, noteworthy patterns emerged. Notably, instances, where the chosen AI teammate outperformed the other AI agent in terms of overall accuracy and recent performance (i.e., the number of correct suggestions in the last 5 tasks), were associated with an increased likelihood of the participants retaining their current AI teammate. Specifically, within scenarios where the participants had the option to switch AI teammates and had prior experience with both agents, the average discrepancies favouring the currently selected AI teammate were 15% (MD = 13%; SD = 21%) for overall accuracy and 1.30 tasks (MD = 1 task; SD = 1.32 tasks) for recent performance. Conversely, when the participants opted to switch AI teammates, these disparities diminished to 2% (MD

Factor	# /20	Example Scenario
Overall AI accuracy	15	P6: selected which AI agent to work with "by overall accuracy"
Recent AI accuracy	13	P20: used the "last five tasks" accuracy to select AI teammates
Patterns	5	P8: used an AI's records to predict its future correctness
Bad runs	3	P16: switched to the other AI after "a few wrong suggestions in a row"
Exploration	20	All participants tried out both AI teammates
Implied order	18	All but 2 participants selected Agent A first
Obvious error (preferred)	5	P18: prefer to receive and reject obviously wrong suggestions
Obvious error (disliked)	5	P4: obvious errors discredited AI agents as "suggesters"
Team accuracy (considered)	6	P11: compared human-AI team performances with different agents
Team accuracy (disregard)	5	P6: a correct suggestion was good regardless of whether I complied with it
Past compliance	4	P18: more likely to accept after incorrectly rejecting a suggestion
Task difficulty	2	P2: AIs cannot be directly compared due to varying task difficulties
Helpfulness for difficult tasks	2	P13: preferred an AI since it was more helpful for difficult tasks
Choice-supportive bias	2	P15: became too "comfortable" to switch after working with the same AI agent for a long period of time
Extrinsic Motivation	2	P18: lost hope to win a performance award, hence switched AI agents purely due to "frustration"
First impression	1	P3: felt "stupid" for selecting an AI for a long time due to its poor performance for its first 5 tasks

Table 4.1: Identified Factors that Affected AI Selection from the Exploration Study

= 0%; SD = 19%) for overall accuracy and 0 tasks (MD = 0; SD = 0.2 tasks) for recent performance. Qualitative findings pertaining to RQ1 are expounded below.

AI agent accuracy deeply influenced the participants' decisions regarding AI teammate selection. Notably, 75% (n=15) of the participants explicitly articulated a preference for the agent exhibiting higher overall accuracy during the semi-structured interviews. A popular AI teammate selection approach among the participants entailed the calculation of the overall accuracy for both agents, defined as the ratio of correct suggestions to the total number of tasks each agent had provided suggestions for. Good overall accuracy may compensate for bad recent performance. P10, for example, "tried to consider [AI teammates' recent performance and since] both A and B [were] bad for the last 5 [tasks], but [chose] Agent B [since it was] better overall. Another example was P9, who still "[had] faith" in an AI teammate and continued to stay with it despite it did poorly in the last 5 tasks. Certain participants adopted nuanced perspectives regarding AI agent accuracy. P7 disclosed a holistic approach, noting that they considered AI agent accuracies "from the beginning and at the end equally." Similarly, P2, stated they would "just think about the [AI accuracies as] sums, and not the position of correct and incorrect suggestions [in the grids]." Moreover, P20 noted that although their AI selections were more influenced by the AI teammates' overall accuracy, these selections were also affected by the participant's own correctness records while working with each of the two AI agents. Additionally, 8 participants demonstrated an awareness of their chosen AI teammate's performance trajectory over the course of tasks, with intentions to switch teammates when the current selected AI agents' performance deteriorated. Among these 15 participants, only 13% (n=2) opted for the agent with lower overall accuracy as frequently as the alternative agent. Furthermore, 13 participants indicated their consideration of the agents' accuracies in relation to their most recent 5 tasks. P20 articulated the view that the "last five tasks" accuracy of the AI agents was more important compared to the overall accuracy. Similarly, P11 said they selected AI teammates "mostly" by comparing their recent performances. Empirical observations largely corroborated these statements, as evidenced by the behaviour of these 13 participants. Specifically, within the 76 AI teammate selection opportunities that arose for these 13 participants, where both AI agents had contributed to at least 5 tasks for each considered participant, and had exhibited different accuracies in the most recent 5 tasks, the participants selected the AI agent that had the higher accuracy in 87% (n=66) of the instances.

Another prevalent strategy entailed the identification of patterns within the AI agents' suggestions and correctness records. P3 articulated that "if the two AI agents were accurate, trying to identify patterns in the AI agents' records would not be required," indicating they tried to identify patterns in the AI agents' performance records due to the

AI agents' unreliability. Several other participants, namely P8, P17, P18, and P19, similarly attempted to identify patterns within the agents' suggestions. These participants anticipated a degree of consistency from the agents and employed identified patterns as a heuristic to anticipate the correctness of forthcoming suggestions. P8 claimed to perceive a recurring pattern wherein an agent "always made 1 incorrect suggestion for every 4 to 5 tasks" Analogously, P18 decided to reject one of the provided suggestions, citing a sequence of "correct 3 times in a row previously" as an indicator that "[the AI agent] would not make 4 correct suggestions in a row." Furthermore, some participants expressed their additional considerations to consecutive incorrect suggestions. For instance, P16 expressed their tendency to switch to the other AI agent when their current teammate exhibited a sequence of "a few wrong suggestions in a row." Parallelly, P7 and P18 opted to switch to the alternative AI agent due to their perception that the current AI teammate had undergone a period of suboptimal performance, characterized by "a bad run" marked by "3 or 4 incorrect suggestions in a row."

All participants undertook proactive efforts to accumulate information concerning both AI agents over the course of the exploration study; however, despite asserting that their initial selection of the first AI agent was random, 90% (n=18) of the participants chose Agent A for their initial set of 5 tasks. Notably, the interface, depicted in Figure 3.7.b, did not suggest a default AI agent. We posit that the button placement and AI agent nomenclature may have implicitly conveyed an order, potentially influencing the participants' AI teammate selections. Moreover, all participants, with the exception of P5, decided to switch to the alternate AI agent for the second set of 5 tasks. P5 elected to engage the second agent for the third block of tasks, driven by a desire to amass "at least 10 tasks for reference for each AI agent." Besides, several participants articulated that they utilized AI teammate switches as a means of gathering additional information about the alternative AI agent. For example, P11's rationale centred on the belief that "the tasks had become more difficult over time," prompting the switch to inspect whether the chosen teammate "could deal with more difficult tasks." Similarly, participants including P16 and P17 opted to switch to the other agent, under the assumption that their current selection had been provided ample opportunities but still failed to "redeem itself." In addition, P2, P5, P6 P7, P9, P10, P13, and P20 explicitly said they wanted to examine how both AI agents perform before settling with an AI teammate.

The participants exhibited contrasting preferences regarding incorrect suggestions when it came to their AI teammate selection. Specifically, considering the case when the suggestions were incorrect, some favoured AI agents that provided incorrect suggestions resembling the correct answers, while others leaned towards AI teammates that offered incorrect suggestions with noticeable distinction from the correct responses. For example, Figure 4.1

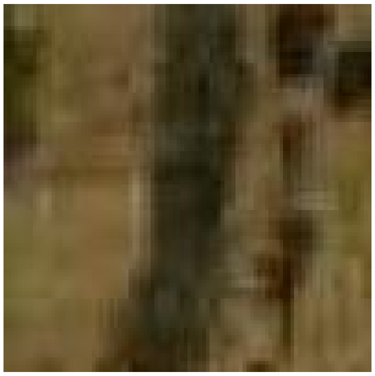


Figure 4.1: Example Annotated Image from the Exploration Study

displays one of the annotated images used for the exploration study. In this case, the correct answer is “I”; an incorrect suggestion noticeably distinct from the correct answer is “Ξ” whereas an incorrect suggestion that resembles the correct answer is “T.” Notably, despite both AI agents' annotation behaviours adhering to identical randomization protocols, participants such as P8, P11, P15, P18, and P20 expressed a propensity for one of the agents known for providing noticeably incorrect suggestions. Their rationale revolved around perceiving such an AI agent as “more useful,” primarily because when confronted with an overtly erroneous suggestion these participants could promptly “simply reject it.” Conversely, participants including P4, P6, P9, P17, and P19 voiced a preference for AI agents offering suggestions that resembled the correct answers. For instance, P4 explained that their preference stemmed from their treatment of both AI agents as mere “suggesters,” indicating that providing obviously inaccurate suggestions would undermine an AI agent's credibility.

The participants exhibited varying attitudes concerning the relevance of their own annotation performance (i.e., the human-AI team accuracy) in the incorrectly rejected selection of their AI teammate. Some participants, namely P9, P11, P13, P15, P17, and P20, actively assessed their performance with each of the two AI agents before making a choice regarding their next AI teammate. For instance, P11 would compare how well the human-AI teams performed overall when each of the two AI teammates was selected. P17, on the other hand, would compare the human-AI team's short-term performances with different AI teammates. Contrarily, an opposing group of participants, including P3, P6, P7, P12, and P16, expressed indifference toward their own annotation performance when collaborating with the AI agents. P6 explained the rationales behind this perspective, emphasizing that “even if [the final answer was] wrong, if the suggestion [was] correct, the AI agent [was] still performing well" and hence, P6 believed the human-AI teams' correctness records did not serve as a reliable indicator AI agents' reliability.

Certain participants, including P2, P15, P17, and P18, stated their AI teammate selection decision-making was influenced by their compliance history. Notably, these participants adopted varied rationales and approaches in this regard. P2 assessed whether they had "previously rejected correct suggestions." Should this be the case, they expressed a heightened level of trust in the AI agent. In contrast, P5 based their decision on the number of suggestions accepted that were deemed "actually helpful." Participants P17 and P18 evaluated the counts of incorrectly accepted suggestions and incorrectly rejected suggestions, respectively. On the other hand, some participants did not consider their compliance history. Conversely, P12 refrained from switching AI teammates multiple times, even after inaccurately accepting or rejecting suggestions on multiple occasions within the previous 5 tasks. According to P12, their criterion for selection predominantly revolved around the correctness of the agents' suggestions, rather than their compliance with the suggestions. Similarly, P7 believed that regardless of their suggestion acceptance decisions, the correctness of these suggestions "didn't play a role in [the AI agents'] abilities."

Furthermore, several factors contributed to the participants' decision to switch AI teammates. These factors were mentioned by a small number of participants:

- *Task difficulty*: P2 believed the AI agents' accuracies should be compared directly since "some [tasks] were more difficult than others," an opinion also shared by P3 and P11.

- *Helpfulness for difficult tasks*: P12 and P13 preferred an agent because it provided correct suggestions when P12 and P13 had "low confidence" in and "no clue" for their provisional answers, respectively.

- *Choice-supportive bias*: P4 and P15 admitted they stayed with an agent due to overestimating the accuracy since they trusted the agent they chose "more and more" and felt too "comfortable" to switch after working with the same AI agent over many tasks, respectively.

- *Extrinsic motivation*: P2 stated that they did not switch to the other AI agent because they felt the current AI agent was "good enough," although not great, and since there were only a few tasks left and they still needed several points to win a performance-based award, they chose to stay with the same AI agent and not take the risk. Conversely, P18 switched AI teammates for the last 5 tasks due to "frustration" after working with the same AI teammate for 30 tasks consecutively, since they had realized they no longer had the possibility to win any performance-based rewards.

Factor	# /20	Example Scenario
Sense-making of suggestions	15	P6: tried to match the annotated images with the provisional answers and the AI-generated suggestions
Prioritizing AI answers	5	P3: likely to accept the suggestions when they did not feel "competent"
Prioritizing human answers	6	P14: "inclined" to choose their own answers over AI suggestions
Matching top options	6	P8: treated suggestions as tie-breakers between two options
Providing new perspectives	2	P9: likely to accept new ideas from AI agents that made sense
Past AI accuracy	13	P7: checked how accurate an AI was when for "50/50" cases to decide whether to accept the suggestions
AI performance for specific letters	4	P4: when Agent A suggested "A" I usually accepted the suggestions
Past incorrect suggestion acceptances & rejections	7	P15: previous incorrect rejections may lead to inclination to accept future suggestions

Table 4.2: Identified Factors that Affected AI Compliance from the Exploration Study

- *First impression*: P3 worked with an AI agent for only 5 tasks, which the agent performed poorly for, then did not give that agent a second chance until the 8th group of tasks. Once the participant switched back to this agent, this agent became more accurate, which happened by chance; regardless, the participant commented that the "first impression [of the AI agent was] deceiving" and they felt "stupid" for not giving that AI agent a second chance earlier.

4.2 Main Factors that Affected AI Teammate Compliance (RQ2)

To explore the factors influencing the participants' compliance with AI agents' suggestions, we conducted both quantitative and qualitative analyses. We amassed a dataset comprising 658 decisions where the participants opted either to accept or reject the selected AI agent's suggestion when it differed from the corresponding participant's provisional answer. After GLMMs and likelihood ratio tests, we identified three significant factors in the participants' decision-making processes. Below, we delve into the quantitative and qualitative results to address RQ2.

The most prominent factor shaping the participants' acceptance of AI agent suggestions was the sense-making of the suggestions. Fifteen participants explicitly mentioned that they would assess whether their provisional answer or the agent's suggestion aligned better with the annotated image. For instance, P6 mentioned that they would "check whether [their provisional answer] or the agent's suggestion [matched the image] more." Eight participants also articulated their inclination to reject suggestions that lacked similarity to the corresponding annotated images. P8, for instance, said they trusted an AI teammate but when it made obvious mistakes they would still reject its suggestions. Given this strong emphasis on sense-making, the participants were significantly more likely to accept the AI agent's suggestion when it was correct, $\beta = 1.73, t(634) = 8.86, p < .001$.

The participants varied in terms of prioritizing either the AI agent's suggestions or their own answers. Thirty percent (n=6) of the participants indicated that they would prioritize the agent's suggestion when uncertain about both their provisional answer and the AI agent's suggestion. For instance, P3 said when they did not feel "competent," they were likely to accept the suggestions. Conversely, 25% (n=5) of the participants mentioned that they would prioritize their provisional answers unless a highly compelling reason favoured the AI agent's suggestion. For example, P14 said they were more "inclined" to choose their own answers although they also considered the suggestions. Our dataset revealed a general tendency among the participants to prioritize their provisional answers. In 685 cases where the participants' provisional answers differed from the agents' suggestions, only 33% (n=196) of suggestions were accepted. Among these 685 cases, 30% (n=207) of suggestions were correct, but merely 53% (n=110) of the correct suggestions were accepted.

The compliance of the participants with their AI teammates' suggestions may hinge on factors including the degree of alignment between the suggestion' and the participants' preferences as well as the novelty and insightfulness of the suggestions. Of the 20 participants, 30% (n=6) mentioned that they would be more inclined to accept an AI agent's suggestion when it closely matched one of their top choices. For instance, P8 described the AI agent as a "tiebreaker," while P9 articulated a scoring system, allocating extra points to the AI agent's suggestion in a weighted selection process. This involved giving each possible answer a score adding up to 100, and to "add 20 [points] extra to the suggestion." Subsequently, they would submit the letter that had the higher score between the agent's suggestion and their own provisional answer. Other participants considered an AI agent's suggestion more favourably if it provided a "new perspective" or "new idea" for interpreting the annotated image. This effect, where participants initially did not regard the suggestion as a potential answer but later found merit in it, was highlighted by participants

P8, P9, and P10.

Among the participants, 65% (n=13) took various aspects of AI agent accuracy into account when deciding whether to accept a suggestion. For instance, P5 sought to calculate the overall accuracy of the AI agent by tallying the number of correct answers it had provided. On the other hand, P8 and P12 monitored the AI agents' accuracies in the "past 5 tasks." Moreover, P7 and P8 considered the chosen AI agent's accuracy primarily when they had approximately"50/50" confidence in their provisional answers and the agent's suggestion. Furthermore, four of these 13 participants factored in the AI agent's performance with specific letters. For example, P4 believed that one of the AI agents was "usually right" when suggesting"*A*." Our collected data also underscores the impact of AI agent accuracy on participant behaviour. Specifically, when the chosen agent's overall suggestion accuracy was higher, the participants were significantly more likely to accept the agent's suggestion, $\beta = 1.13, t(634) = 3.09, p = .002$. However, two participants (P1 and P6) expressed indifference to the accuracy of the AI agent when deciding whether to accept its suggestion. P1 attributed this to the nature of past tasks being "different than the current one," while P6 prioritized the "closest match" between the annotated image and the provisional answer or suggestion.

Another factor influencing AI teammate compliance was the participants' previous decisions regarding the AI agents' suggestions. Seven participants indicated that they would consider whether they had recently accepted incorrect suggestions or rejected correct ones while making subsequent AI suggestion acceptance decisions. For instance, P15 mentioned that they would assess whether they had "rejected a correct suggestion" recently, which would make them more likely to accept the current suggestion. According to our GLMM analysis, when the participants had accepted an incorrect suggestion for the previous task from the chosen agent, they were less likely to accept the suggestion for the current task, $\beta = -0.82, t(634) = -2.46, p = .01$.

4.3 Reactions to Different Performance Scenarios

During the interview, we gauged the participants' reactions to five scenarios in how the human-AI team performed in the task:

1. When both the participant and the AI agent were correct.
2. When the participant rejected the chosen AI agent's suggestion (that they deemed to be incorrect) and the participant's provisional answer was correct.

3. When the AI agent's suggestion was correct but the participant rejected the correct suggestion and submitted their own incorrect provisional answer as the final answer.
4. When both the AI agent's suggestion and the participant's final answer were incorrect and they were different.
5. When both the AI agent's suggestion and the participant's final answer were incorrect and they were identical.

Specifically, we investigated the participants' reactions to various types of human-AI correctness scenarios, such as the one shown in Figure 3.8, which revealed distinct patterns of trust, validation, and frustration. We discuss these scenarios in detail below.

In the first scenario, the participants often reacted by placing more trust in their chosen AI teammate. This increased trust that demonstrated in multiple ways, such as a greater likelihood of staying with the same AI agent in the future, as noted by P14. For P19, it meant being "more willing to accept the AI agent's tasks in the following [tasks]." The participants in this category also reported feeling validated and more confident in their decision-making abilities. However, not all aspects of this scenario were universally liked. P13, for instance, mentioned feeling like they "took the suggestion" even when they had independently arrived at the answer. Similarly, P17 perceived the AI agent as unhelpful, merely replicating their provisional answer, albeit correctly. In contrast, P15 appreciated the simplicity of not having to choose whether to accept the suggestion, and P20 regarded this scenario as "expected," given their anticipation that AI agents should consistently outperform humans.

Conversely, the participants who had correct provisional answers but rejected incorrect suggestions expressed decreased trust in the AI agent. P16 contemplated switching to the other AI agent, especially if this pattern occurred multiple times "in a row." P14 reported feeling "frustrated" and more inclined to trust their own answers in subsequent tasks. P15 differentiated between task difficulty levels, perceiving the AI agent's attempts to "confuse" in easy tasks. P18 and P19 had similar responses, with both participants finding the situation "okay" while both acknowledging the unreliability of the AI agent. P13, on the other hand, felt "really good" about outperforming the agent.

Multiple participants stated they trusted the AI agent more when they incorrectly rejected a correct suggestion; however, P19 remained cautious and continued to compare suggestions with their provisional answers. Meanwhile, some participants saw this outcome as expected due to their belief that AI agents should outperform humans. Conversely, others felt the AI agent's suggestion still appeared incorrect, with P14 becoming "skeptical" of *MIA*'s credibility for judging final answers.

Sce. #	Reaction
1	greater likelihood of staying with the same AI agent
1	greater likelihood of accepting suggestions from the AI agent
1	feeling validated
1	feeling the AI agent replicated their work
1	feeling good about not having to choose whether to accept the answer
1	feeling the scenario is expected as AI agents should outperform humans
2	losing trust in the AI agent
2	feeling frustrated
2	feeling okay due to low expectation in AI accuracy
2	feeling good to outperform the AI agent
3	increasing trust in the AI agent
3	remaining cautious for future tasks regarding AI suggestions
3	feeling this scenario is expected as AI agents should outperform humans
3	feeling *MIA* incorrectly judged final answers
4	feeling accomplished for not being inferior to the AI agent
4	feeling the task might be too difficult for the AI agent
4	decreasing trust in the AI agent
5	feeling this scenario was a shared learning experience between the AI agent and themself
5	feeling the AI agent made a good prediction and showed its ability was not inferior to humans
5	distrusting *MIA*'s ability in judging answers
5	feeling the AI agent intentionally obstructed the participant by providing incorrect suggestions
5	feeling shocked for having matching but incorrect answers
5	feeling this scenario was expected due to anticipating AI agent not to outperform humans

Table 4.3: Identified Participant Reactions to Various Human-AI Performance Scenarios from the Exploration Study

In the scenario where the participant's provisional answer and the AI's suggestion differed and both were incorrect, the participants commonly experienced a sense of achievement of not feeling inferior to their AI teammates. P15 noted feeling "better" in these cases, seeing aspects in the image that even the AI agent had missed. P19 perceived the AI agents as "bad," while P18 considered the tasks "probably too difficult." P17 believed the AI agent should have known the answer, labelling it as "dumb," which reduced trust in the chosen AI teammate.

In the last scenario where both the provisional answer and the AI agent's suggestion were identical but incorrect, the participants, in general, did not exhibit decreased trust in their chosen AI teammate. Instead, several distinct reactions emerged within this context. Some participants viewed this scenario as a shared learning experience, emphasizing that neither the chosen AI agents nor themselves were infallible. For instance, P13 expressed feeling that both the agent and themselves were "not perfect" and that they were both "still learning." Similarly, P20 regarded the AI agent's performance as making a "good prediction," perceiving it as not significantly worse than a human counterpart. On the other hand, some participants held the system responsible for grading answers incorrectly, absolving the AI agent of blame. This perspective reflected a belief that the AI agent had performed adequately, and that any shortcomings were attributed to external factors. Conversely, other participants expected better performance from the AI agent in this scenario and viewed its actions as intentional attempts to obstruct their attainment of performance-based rewards. P17, for example, suspected the AI agent of trying to "keep [them] from" achieving these rewards. There was a range of emotional responses, with some participants "shocked" by the mutual incorrectness, while others perceived the situation as entirely expected. P19, for instance, stated that they did not "expect the [selected AI agent] to do better than [them]," leading them to anticipate incorrect answers despite the match between their provisional answer and the AI agent's suggestion.

4.4 Noteworthy Miscellaneous Quotations

The participants shared miscellaneous perceptions of the AI agents during their engagement in the annotation tasks and interviews. While these perceptions may not directly influence the participants' decisions during tasks, they shed light on how individuals interpret AI agents based on their design and behaviours. These insights provide valuable information for understanding human interactions with AI systems:

- Some participants expected the AI agents to learn and improve performance as they worked on more tasks.

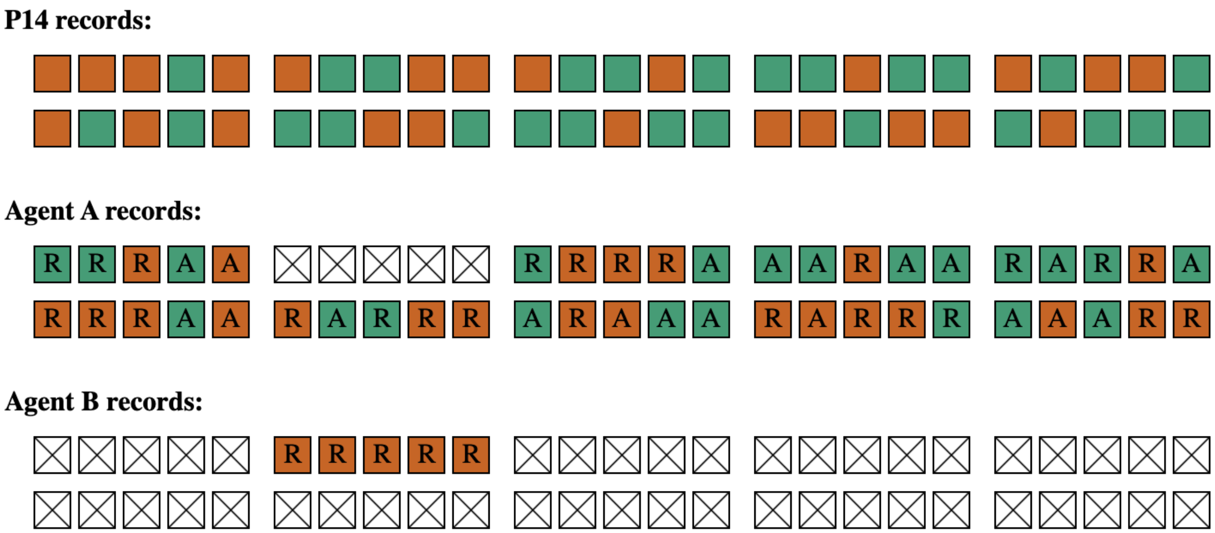


Figure 4.2: Summarized P14 and Their AI Teammates' Annotation Records from the Exploration Study

- Most participants made up narratives to explain the observed behavioural differences between the two AI teammates.

To elaborate, a common impression among the participants was that the AI agents were continuously learning as they engaged in more tasks. For instance, P14, as shown in Figure 4.2, chose to collaborate with Agent A for 45 out of 50 tasks and tried Agent B only once, believing that continued collaboration would allow them to "train" Agent A, leading to improved performance. Similarly, both P13 and P16 perceived their chosen AI teammates as learners and were not disheartened when these AI teammates made mistakes. As a result, they switched AI teammates only three times out of nine opportunities.

Another notable finding is that 18 participants (90%) constructed narratives to explain the distinctions between the two AI agents. For example, P11 selected Agent B more frequently, noting that Agent B was "more predictable" compared to Agent A. Similarly, P13 felt that Agent B consistently "helped [them] out" when they struggled to choose between two answers and that they "worked better together;" in contrast, Agent A was less trusted by this participant. There were two exceptions to these narratives: P20 expressed a lack of trust in both AI agents, attributing their choice to the agents "not [performing] very well," while P16 stated that both AI agents "gave right and wrong answers." Interestingly,

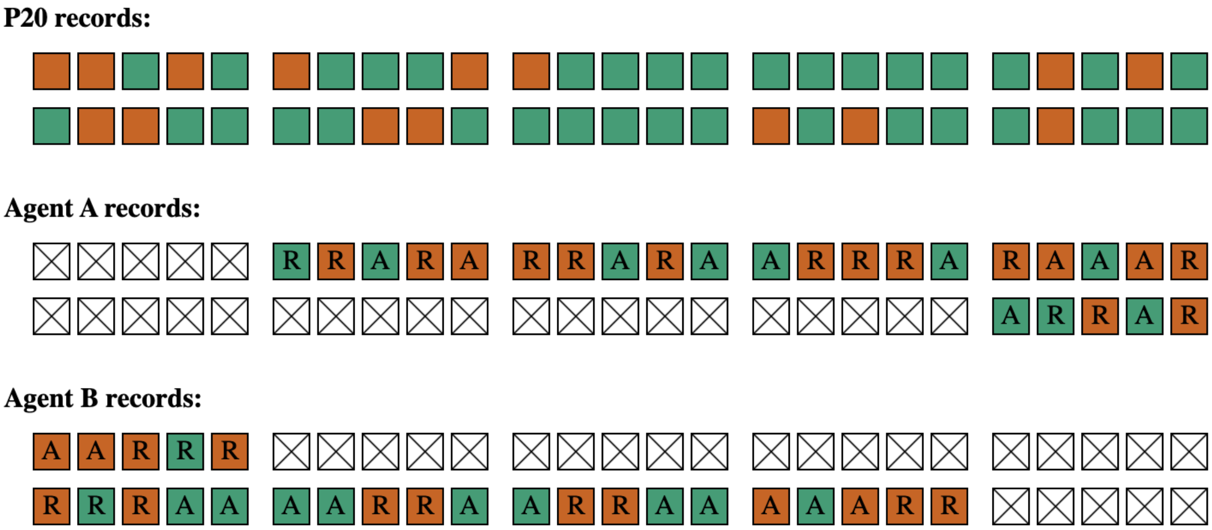

P20 records:
Agent A records:
R R A R A R R A R A A R R R A R A A A R
A R R A R
Agent B records:
A A R R R
R R R A A A A R R A A R R A A A A A R R

Figure 4.3: Summarized P20 and Their AI Teammates' Annotation Records from the Exploration Study

the annotation behaviour of these two participants supported their comments; as shown in Figure 4.3, P20 distributed the 50 tasks evenly between the two AI agents; P16, similarly, distributed at least 20 tasks to each AI agent.

Chapter 5

Discussion

5.1 Design Implications

In this section, we discuss how the results of both the exploration study and the feasibility study inform potential changes to improve *MIA*. Particularly, we will first discuss the existing weaknesses in *MIA*'s current designs. Then, we will move on to explore a potential new design of *MIA*, which has the aims to rectify weaknesses of the existing designs of *MIA* as well as to improve *MIA*'s adaptability and scalability.

5.1.1 Existing Designs of *MIA*

As illustrated in the Methodology section, *MIA* currently has two interface designs. Particularly, *MIA*'s interface design from the exploration study (Figures 3.3 and 3.4) is for image annotation tasks where the images contain exactly 1 letter to annotate, and human annotators are responsible for choosing an AI teammate from two options to receive suggestions from as well as deciding the final answer for each task; whereas the *MIA* interface design from the feasibility study (Figure 3.5) enables human and one AI teammate to work on the same image simultaneously by making multiple annotations with 2-dimensional locations. Below, we will discuss the potential improvable aspects of *MIA*'s designs from both the exploration study and the feasibility study based on the results from these two studies.

***MIA*'s Interface Design from the Exploration Study**

According to the results of the exploration study, the current interface design of *MIA* has the following weaknesses:

- causes annotator assumptions and biases;
- provides correctness feedback too frequently;
- offers little information for assisting suggestion compliance decisions;
- utilizes a decision-making workflow that limits human-AI information exchange.

Each aforementioned weakness will be discussed in detail below. Note that since the design of the *MIA* interface from the exploration study is not designed to have maximized effectiveness as an annotation interface; instead, its primary goal is facilitating an exploratory study by avoiding over-guidance on the participants caused by revealing excessive information about the AI agents and their decisions. Therefore, the aforementioned weaknesses do not indicate inadequacy in the design of the exploration study.

Firstly, the lack of information regarding the AI teammates' characteristics forced the participants to make and rely on false assumptions about their AI teammates to predict their future behaviours. These assumptions affected the participants' AI teammate selection and compliance processes. For instance, some participants would prefer to work with an AI teammate because they assumed it had more consistent performance than the other AI teammate option based on limited observation. Relevantly, we discovered that the vast majority of the participants believed their AI teammate selection impacted the quality of the suggestions they received, while multiple participants have convinced themselves that one AI teammate was superior compared to the other option, although both AI teammates' behaviours were randomized using the same protocol. Since human annotators may make false assumptions about their AI teammates based on limited evidence, in a real-world scenario, especially when the behavioural differences between AI teammate options are difficult to notice, the provision of information regarding AI teammate characteristics may prevent human annotators from basing their AI teammate selection and compliance decisions on false assumptions. Notably, many findings from the field of explainable AI may facilitate such information provision.

Secondly, the correctness feedback timing is fixed. The results of the exploration study, especially the reactions of the participants to the pop-up windows, which appeared after the completion of each task and showed whether the corresponding participant and the selected AI agent were correct, show that the correctness feedback had instant impacts on the participants' perceptions of themselves and their AI teammates. This illustrates a potential weakness of the current interface design of *MIA*. Particularly, this instant impact on humans' perceptions may not be ideal in many cases since the annotators' decisions may be overly influenced by the most recent, instead of long-term, performances of themselves and the AI teammates.

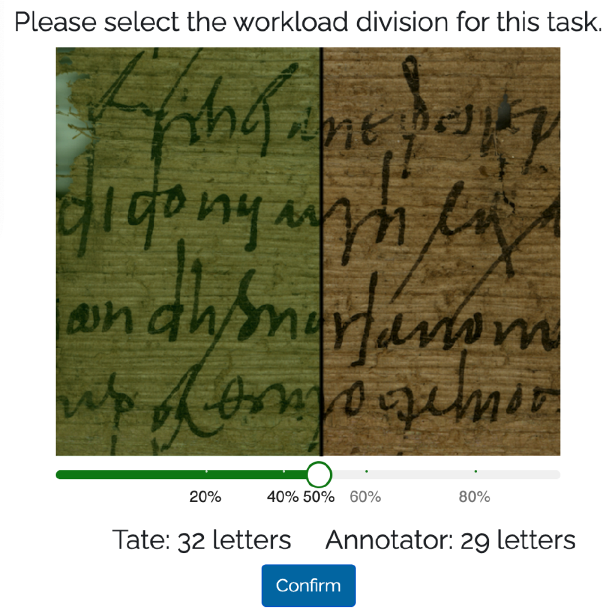
Thirdly, explanations of the AI teammates' decisions are not provided. The participants of the exploration study generally belonged to one of two categories when they could not decide whether to accept an AI suggestion—one category of the participants would prioritize their own provisional answer, while the participants in the other category would prioritize the received AI suggestion. This phenomenon could be a result of the lack of information regarding the explanation of the AI teammates' decisions.

Fourthly, the current workflow could be modified to allow human annotators and their AI teammates to exchange more information. The current workflow of *MIA* requires the annotator to input exactly one provisional answer before receiving a suggestion from the selected AI teammate. According to the results of the exploration study, the participants usually accepted suggestions when they aligned with their top choices or provided novel perspectives. This information can be used to design new workflows for more effective human-AI decision-making processes.

***MIA*'s Interface Design from the Feasibility Study**

A feasibility study deployed with *MIA*'s alternate interface design has previously been discussed in [202]; however, only a small portion of the feasibility study results are reported in [202]. Hence, most of the discussed discovered weaknesses of *MIA*'s alternate interface will be based on results that are not included in [202]. Below, we will outline the design of the feasibility study; then, we will discuss how the results of this study inform further development of *MIA*.

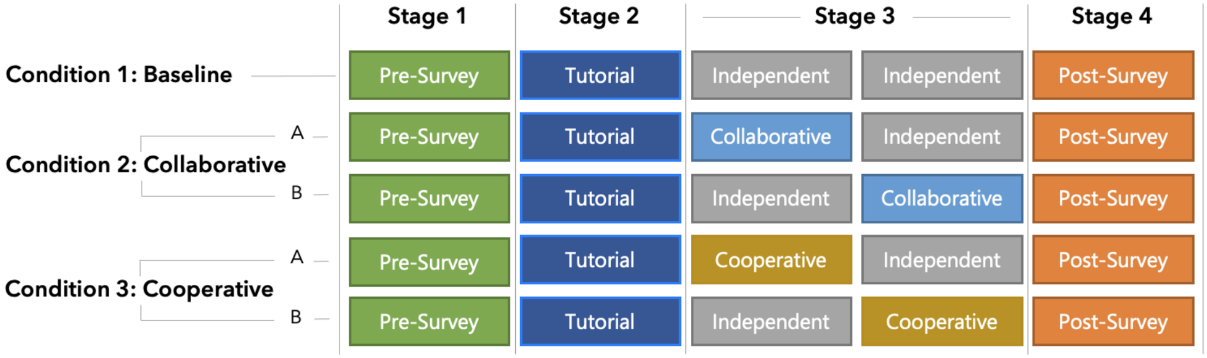
The procedure of the feasibility study is shown in Figure 5.2. Each participant in the feasibility study was assigned to one of the five conditions (i.e., conditions 1, 2A, 2B, 3A, 3B) shown in Figure 5.2 to annotate multiple papyrus images. Before starting the annotation tasks, each participant completed a questionnaire about their demographic information and watched a 3-minute tutorial video to familiarize themself with how to use the given annotation interface (Figure 3.5); note that the tutorial video differed slightly for different conditions since some actions could be taken only by the participants in certain conditions (e.g., seeking help from the AI agent). Then, for each image, the corresponding



Please select the workload division for this task.

Tate: 32 letters Annotator: 29 letters

Figure 5.1: Pop-Up Window for Image Division from the Feasibility Study



		Stage 1	Stage 2	Stage 3		Stage 4
Condition 1: Baseline		Pre-Survey	Tutorial	Independent	Independent	Post-Survey
Condition 2: Collaborative	A	Pre-Survey	Tutorial	Collaborative	Independent	Post-Survey
	B	Pre-Survey	Tutorial	Independent	Collaborative	Post-Survey
Condition 3: Cooperative	A	Pre-Survey	Tutorial	Cooperative	Independent	Post-Survey
	B	Pre-Survey	Tutorial	Independent	Cooperative	Post-Survey

Figure 5.2: Procedure of the Feasibility Study

participant either annotated alone (i.e., the “Independent” box) or alongside an assigned AI agent; when a participant annotated alongside an AI agent, the AI agent either only annotated by itself without interacting with the participant (i.e., the “collaborative” box) or both sought help from and gave help to the participant (i.e., the “cooperative” box).

Before starting annotating each image, when the participant annotated with an AI agent, the participant would be asked to choose how to divide the workload for the assigned image between the AI agent and themself using the pop-up window shown in Figure 5.1. While annotating an image alone or when the AI agent was present but did not interact with the participant, the participant's sole goal for the task was to locate and label all letters in their assigned area of the image; to do so, they would need to click on where each letter was located on the image, then input the corresponding label (i.e., letter) using a virtual keyboard. When an AI agent was present and interacted with the participant, the participant could flag any of the annotations they previously made to seek a suggestion; note that the AI agent may refuse to provide help to some flagged annotations. Similarly, the AI agent may seek help by flagging its annotations, in which case the participants may opt to provide their suggestions to the AI agent. The AI agent took 1 to 5 seconds to create each annotation/suggestion; the accuracy for the AI agent's own annotations was set to 85% and for its suggestions was 75%. Similar to the AI agents in the exploration study, the feasibility study's AI agent was also aware of the ground truths and would make incorrect outputs based on the Greek Letter Oriented Substitution Matrix from [214]. In addition, when an AI agent was present, the participant could hover their cursor on the AI agent's annotations to view the AI agent's labels, regardless of whether the AI agent interacted with the participant; this design was included to enable the participants' observational learning. Moreover, each participant also completed pop-up surveys during the annotation tasks, which investigated their perceptions of the tasks and the AI agent (e.g., the participants' self-report engagement levels). In total, 51 participants took part in the feasibility study and they were remunerated with both a base payment and a performance-based bonus, which correlated to the number of annotations, not including suggestions, they made as well as these annotations' accuracy. After completing the annotation tasks, each participant also completed a questionnaire that included a Self-Report Altruism scale [180] and investigated their experience during the annotation tasks (e.g., whether they enjoyed cooperating with the AI agent).

The alternate *MIA* interface design has three main weaknesses based on the feasibility study results:

- encourages help-seeking behaviour insufficiently;
- fails to clearly show that AI help requests are optional to fulfill;
- help-giving interaction slows down annotation speed.

Below, each weakness will be discussed in detail. Notably, unlike the exploration study, one of the above weaknesses negatively affected the effectiveness of the feasibility study.

Particularly, the lack of help-seeking behaviour was not expected; consequently, little data was collected for relevant analysis.

Firstly, while working alongside an AI agent and given the opportunity to interact with it for help-seeking and help-giving, the participants generally sought help infrequently. As a result, the annotation accuracy of the participants when they could seek help was not significantly higher than when they worked alongside an AI agent without the opportunity to do so. Since a performance-based bonus existed and all participants had learned how to use the annotation interface by watching a tutorial video, the lack of help-seeking behaviour was most likely due to the interface design. Specifically, the participants needed to flag an annotation to seek help with it by clicking the flag button located in the corresponding annotation's box in the annotation panel; then, it took approximately 1-5 seconds for the AI agent to provide its corresponding suggestion, if it decided to do so. Presumably, the interface design failed to sufficiently encourage help-seeking behaviour from the participants due to the lack of positive feedback when the participants did so.

Secondly, the interface over-encourages help-giving behaviour. The feasibility study's results showed that despite the participants' self-report altruism scores correlating positively with their self-report satisfaction levels for assisting the AI agent, these data do not correlate with the frequency of the participants' help-giving behaviour. This is mainly because most participants frequently provided help to the AI agent while asked to. Presumably, most participants interpreted the help requests from the AI agent as orders and felt obligated to provide suggestions; while most participants fulfilled the AI agents' help requests, the participants who were more altruistic could be more willing to do so. This could be a weakness of the alternate design of *MIA*; as although the interface successfully encouraged help-giving behaviour, it may be ideal that the users do not provide assistance to the AI when they are not certain about theirs, since incorrect suggestions may harm the AI agent's performance.

Thirdly, the feasibility study results show that the participants annotated significantly slower while they could seek help from and give help to the AI agent, compared to when the AI agent was present but did not interact with the participants. Since most participants rarely sought help and gave help frequently, this difference in speed in different conditions was most likely caused by the participants' help-seeking behaviour. Therefore, using the data collected through the feasibility study, we conclude that another existing weakness of *MIA*'s alternate interface is it does not allow the human to perform help-giving behaviour in a time-efficient manner.

5.1.2 New Design of *MIA*

Next, we propose a new design of *MIA*. This design is based on not only the discovered weaknesses of *MIA*'s current and alternate interface designs, but also other considerations for improving *MIA*'s adaptability and scalability as a mixed-initiative annotation platform.

Firstly, the AI teammates' characteristics should be displayed to the annotators. The goal of such information is to prevent the annotators from making false assumptions about the AI's capabilities, which could be crucial in assisting the annotators in making rational decisions for AI teammate selection. Some examples of information that could be displayed to avoid false assumptions about the AI agents include whether the AI agent is learning as it performs more tasks, whether the AI agent has low performance for specific types of tasks, the accuracy of the AI agent while working on the validation data set.

Secondly, *MIA* should support task creators to adjust when, whether, and what type of correctness feedback is provided to the annotators. As discussed, delayed feedback may be preferred over immediate feedback such that the annotators' decisions are less affected by the recent performance of their AI teammates. Also, for some tasks, the task creator may prefer not to provide correctness feedback at all, such as when such ground truths are unavailable. Moreover, the task creators may prefer to generate correctness feedback only for certain tasks with ground truths or using estimations (e.g., by inputting the annotators' final answers into a large language model to estimate the quality of those answers) instead of ground truths.

Thirdly, the rationale behind each AI suggestion should be explained to the annotators. Particularly, if the participants had access to explanations regarding why their AI teammates selected their answers, such as by identifying certain features in the assigned image, the participants may use this additional information to decide whether to comply with the AI suggestions through reasoning. Notably, different types of tasks (e.g., image and audio annotation tasks) may require different media for the illustration of the AI suggestion rationales. Hence, *MIA*'s new interface should include a component specifically designed for AI agents to explain their rationales and this component should support media formats (e.g., text, image, audio, video, and interactive web element).

Fourthly, to improve *MIA*'s scalability, built-in adjustable task assignment and AI teammate recommendation algorithms should be incorporated. *MIA* should be able to automatically group human annotators based on their expertise or other characteristics and allow task creators to manually move annotators between groups. Also, the task creators should be able to use the task assignment algorithm to decide which annotator groups (e.g., laypeople vs. experts) and AI agents are assigned which tasks, as well as the number of annotators from each group should complete each task assigned to that group. These features would allow task creators to effectively assign tasks to a large number of potential annotators, whose answers could be verified through majority votes while minimizing the cost of task deployment. *MIA* should also allow task annotators to deploy multiple AI agents simultaneously and choose the number of suggestions the annotators can receive for each task. Moreover, *MIA* should allow task creators to manually assign AI agents to different tasks, as well as input their opinions regarding how much they recommend each AI agent for each type of task. In addition, *MIA* may automatically recommend AI agents to annotators for different types of tasks if the task creator inputs the accuracy data of each agent for each type of task. The adaptation of these features would allow the annotators to choose AI agents more effectively, especially when a large number of potential AI teammates are available. Notably, receiving multiple suggestions from different AI agents for the same task may cause confusion for the annotators; however, it may also allow the annotators to receive more diverse suggestions, which could improve the performance of the human-AI team. We leave the exploration of this configuration to future studies.

Fifthly, to improve *MIA*'s adaptability, multiple types of annotation tasks as well as human-AI help-seeking and help-giving interactions should be supported. Particularly, *MIA* should allow task creators to easily create new tasks by uploading files with different formats, including image, audio, text, and video. *MIA* should be able to turn these uploaded files into annotation tasks; for different file formats, *MIA* should also allow task creators to select what kind of task should be generated; for instance, for each image, task creators should be able to select whether they want annotators to make a single annotation without location, one or more annotation(s) without location(s), or perform image segmentation. Besides, *MIA* should allow task creators to select whether the annotators interact with AI agents with buttons, similar to the two *MIA* described previously, or through chat with natural language and media files, since large language models are becoming increasingly popular.

Lastly, a new human-AI decision-making workflow should be adopted. According to the results of the two aforementioned studies, the participants of the exploration study usually accepted suggestions when they aligned with their top choices or provided novel perspectives; also, the feasibility study participants lacked help-seeking behaviour, overly performed help-giving behaviour, and were slowed down by slow help-giving interactions. Therefore, based on the workflow used in the exploration study, a potential change to the annotator help-seeking workflow could be enabling the annotator to first select one or more (e.g., up to three) top choices; then, the selected AI teammate generates its suggestions by selecting one suggestion from the submitted top options and another suggestion that is not one of the annotator's top options. The AI agent may provide each of these two suggestions only when the AI agent's confidence level surpasses a pre-defined threshold. The advantages of adopting this workflow include allowing the annotators to reconsider options that they may otherwise overlook, ensuring the annotators put in sufficient effort to select their provisional answers, and enabling the AI agents to utilize the information regarding the annotators' confidence levels in different answers to provide suggestions. Notably, this workflow should allow the annotator to proceed by submitting only one top choice to prevent them from submitting less confident answers as top choices and wasting time doing so. To avoid the lack of help-seeking behaviour, this help-seeking workflow may be initiated automatically once the task is entered for tasks where only one annotation is made. For tasks where multiple annotations are created, this workflow can be initiated automatically once an annotation has been created by the annotator. The annotator help-giving workflow should be initiated only when the annotator has high confidence in their suggestion and should be more time-efficient. To achieve these goals, the annotation interface should notify the annotators to provide suggestions only when they are certain about their suggestions, such as by showing a pop-up window with this message the first time an annotator provides a suggestion. Human annotators should be able to provide suggestions to all AI annotations but the AI agents, especially when implemented as active learners, should seek help actively only when their confidence level is below a pre-defined low threshold. Although it is non-trivial to enable AI agents to learn through observing human annotator answers and suggestions, such implementation may be tremendously beneficial as it may allow AI agents to eventually become independent task performers. In addition, when an AI agent seeks help, the annotator should be able to view all relevant information about the help request (e.g., the portion of image that requires help, the AI agent's top choices, the reasoning behind the AI agent's top choices, etc.) in a summarized form through a pop-up banner that does not block the view of the main interface, such that the annotator can carry out their annotation without being overly interrupted by the AI agent help requests; these banners should also automatically pile together in a stack if multiple of them are present simultaneously. The annotator should also be able to quickly decide whether to accept or dismiss each help request and the corresponding banner (e.g., by pressing the space button on the keyboard to accept the help request and any other key to dismiss); once an annotator accepts a help request, the interface should automatically show the input component (e.g., virtual keyboard or text box) and the annotator should be able to immediately input their suggestion without any other actions (e.g., clicks).

5.2 Limitations and Future Work

Both the exploration study (n=20) and the feasibility study (n=51) are limited in terms of sample size and participant demographics, which may constrain the generalizability of our findings. Also, due to the small sample sizes, it was difficult to conduct quantitative analysis with the collected data, such as for identifying correlations between various human annotator behaviours. Hence, future research with larger and more diverse participant groups is needed to validate the identified factors and the proposed new design of *MIA* more comprehensively.

Additionally, the two aforementioned studies are also limited with regard to task configurations; therefore, future studies should investigate the applicability of the discovered factors and the proposed new design of *MIA* across a wider range of mixed-initiative annotation scenarios. Particularly, based on related work, we propose future studies to investigate more task types (e.g., audio labeling [69], video annotation [96], and image segmentation [208]), additional human-AI team configurations (e.g., human(s) assisting an AI agent [88] and human(s) receiving suggestions from multiple AI agents that interact with each other [2]), as well as more diverse feedback types (e.g., feedback produced through natural language generation [31], automatic generated feedback for questions without standardized solutions [103, 76], and delayed feedback [10, 119]). By following these proposed directions, future studies could contribute to a more nuanced understanding of the factors influencing collective human-AI decision-making and provide valuable insights for designing effective human-in-the-loop systems.

Chapter 6

Conclusion

This thesis investigated the various factors influencing the decision-making processes of human members in human-AI teams, particularly concerning AI teammate selection and compliance in the collaborative annotation context. Specifically, our findings from a human-subject think-aloud study shed light on how factors like the AI agents' overall and recent performance as well as AI agents' performance predictability could sway human members' preferences for collaborating with particular agents. Furthermore, we revealed that the alignment of agents' suggestions with human members' preferences and the agents' ability to offer novel perspectives impact the likelihood of compliance with these suggestions. Additionally, our investigation uncovered that human members may construct narratives to interpret variances in AI teammates' behaviour, often relying on limited evidence. This thesis also contributes to human-AI collaborative annotation research by proffering *MIA*, a versatile web platform for mixed-initiative annotation. Based on the results of two mixed-initiative annotation studies performed on *MIA*, this thesis proposes potential changes to *MIA*'s current designs with aims of improving its effectiveness, scalability, and adaptability; notably, these design changes may also be applicable to other mixed-initiative annotation platforms. Lastly, we acknowledged this thesis' limitations related to sample sizes, participant demographics, and task configurations; hence, we proposed avenues for future research to expand and refine our understanding of human-AI collaborative annotation in more extensive and diverse contexts.

References

[1] Jennifer L Aaker and Satoshi Akutsu. Why do people give? the role of identity in giving. *Journal of consumer psychology*, 19(3):267–270, 2009.

[2] Abdelkader Adla, Bakhta Nachet, and Abdelkader Ould-Mahraz. Multi-agents model for web-based collaborative decision support systems. In *ICWIT*, pages 294–299, 2012.

[3] Eirikur Agustsson, Jasper RR Uijlings, and Vittorio Ferrari. Interactive full image segmentation by considering all regions jointly. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11622–11631, 2019.

[4] Nesreen K Ahmed, Amir F Atiya, Neamat El Gayar, and Hisham El-Shishiny. An empirical comparison of machine learning models for time series forecasting. *Econometric reviews*, 29(5-6):594–621, 2010.

[5] Nosayba Al-Azzam and Ibrahem Shatnawi. Comparing supervised and semi-supervised machine learning models on diagnosing breast cancer. *Annals of Medicine and Surgery*, 62:53–64, 2021.

[6] Vincent Aleven, Elmar Stahl, Silke Schworm, Frank Fischer, and Raven Wallace. Help seeking and help design in interactive learning environments. *Review of educational research*, 73(3):277–320, 2003.

[7] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[8] James E Allen, Curry I Guinn, and Eric Horvtz. Mixed-initiative interaction. *IEEE Intelligent Systems and their Applications*, 14(5):14–23, 1999.

[9] Abdullah Almaatouq, Peter Krafft, Yarrow Dunham, David G Rand, and Alex Pentland. Turkers of the world unite: Multilevel in-group bias among crowdworkers on amazon mechanical turk. *Social Psychological and Personality Science*, 11(2):151–159, 2020.

[10] Abdulaziz Derwesh A Almalki and Abdellah Ibrahim Mohammed Elfeky. The effect of immediate and delayed feedback in virtual classes on mathematics students' higher order thinking skills. *Journal of Positive School Psychology*, pages 432–440, 2022.

[11] Amazon. Amazon sagemaker data labeling, 2022.

[12] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. Guidelines for human-ai interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–13, 2019.

[13] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. The role of shared mental models in human-ai teams: a theoretical review. *Theoretical Issues in Ergonomics Science*, 24(2):129–175, 2023.

[14] Zahra Ashktorab, Michael Desmond, Josh Andres, Michael Muller, Narendra Nath Joshi, Michelle Brachman, Aabhas Sharma, Kristina Brimijoin, Qian Pan, Christine T Wolf, et al. Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–27, 2021.

[15] Yannis M Assael, Brendan Shillingford, Shimon Whiteson, and Nando De Freitas. Lipnet: End-to-end sentence-level lipreading. *arXiv preprint arXiv:1611.01599*, 2016.

[16] Nikola Banovic, Zhuoran Yang, Aditya Ramesh, and Alice Liu. Being trustworthy is not enough: How untrustworthy artificial intelligence (ai) can deceive the end-users and gain their trust. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW1):1–17, 2023.

[17] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2429–2437, 2019.

[18] Tobias Bartholomé, Elmar Stahl, Stephanie Pieschl, and Rainer Bromme. What matters in help-seeking? a study of help effectiveness and learner-related factors. *Computers in Human Behavior*, 22(1):113–129, 2006.

[19] C Daniel Batson, Michelle H Bolen, Julie A Cross, and Helen E Neuringer-Benefiel. Where is the altruism in the altruistic personality? *Journal of personality and social psychology*, 50(1):212, 1986.

[20] Gabriele Bellucci, Julia A Camilleri, Simon B Eickhoff, and Frank Krueger. Neural signatures of prosocial behaviors. *Neuroscience & Biobehavioral Reviews*, 118:186–195, 2020.

[21] Raquel Benbunan-Fich and JB Arbaugh. Separating the effects of knowledge construction and group collaboration in learning outcomes of web-based courses. *Information & management*, 43(6):778–793, 2006.

[22] Jeffrey P Bigham, Kristin Williams, Nila Banerjee, and John Zimmerman. Scopist: building a skill ladder into crowd transcription. In *Proceedings of the 14th International Web for All Conference*, pages 1–10, 2017.

[23] Francesco Biondi, Ignacio Alvarez, and Kyeong-Ah Jeong. Human–vehicle cooperation in automated driving: A multidisciplinary review and appraisal. *International Journal of Human–Computer Interaction*, 35(11):932–946, 2019.

[24] Gautam Biswas, Krittaya Leelawong, Daniel Schwartz, Nancy Vye, and The Teachable Agents Group at Vanderbilt. Learning by teaching: A new agent paradigm for educational software. *Applied Artificial Intelligence*, 19(3-4):363–392, 2005.

[25] Kristen Blair, Daniel L Schwartz, Gautam Biswas, and Krittaya Leelawong. Pedagogical agents for learning by teaching: Teachable agents. *Educational Technology*, pages 56–61, 2007.

[26] Samuel Boobier, Anne Osbourn, and John BO Mitchell. Can human experts predict solubility better than computers? *Journal of cheminformatics*, 9:1–14, 2017.

[27] Matthew M Botvinick and Zev B Rosen. Anticipation of cognitive demand during decision-making. *Psychological Research PRPF*, 73:835–842, 2009.

[28] Matthias Brand, Kirsten Labudda, and Hans J Markowitsch. Neuropsychological correlates of decision-making in ambiguous and risky situations. *Neural Networks*, 19(8):1266–1276, 2006.

[29] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: cognitive forcing functions can reduce overreliance on ai in ai-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[30] Marissa Burgermaster, Krzysztof Z Gajos, Patricia Davidson, and Lena Mamykina. The role of explanations in casual observational learning about nutrition. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 4097–4145, 2017.

[31] Anderson Pinheiro Cavalcanti, Arthur Barbosa, Ruan Carvalho, Fred Freitas, Yi-Shan Tsai, Dragan Gašević, and Rafael Ferreira Mello. Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence*, 2:100027, 2021.

[32] Joel Chan, Steven Dang, and Steven P Dow. Ideagens: enabling expert facilitation of crowd brainstorming. In *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, pages 13–16, 2016.

[33] Balasubramaniyan Chandrasekaran and James M Conrad. Human-robot collaboration: A survey. In *SoutheastCon 2015*, pages 1–8. IEEE, 2015.

[34] Catherine C Chase, Doris B Chin, Marily A Oppezzo, and Daniel L Schwartz. Teachable agents and the protégé effect: Increasing the effort towards learning. *Journal of Science Education and Technology*, 18(4):334–352, 2009.

[35] Loh CheeWyai, WaiShiang Cheah, Adib Kabir Chowdhury, and Christian Gulden. Engineering sustainable software: A case study from offline computer support collaborative annotation system. In *2015 9th Malaysian Software Engineering Conference (MySEC)*, pages 272–277. IEEE, 2015.

[36] Jim X Chen. The evolution of computing: Alphago. *Computing in Science & Engineering*, 18(4):4–7, 2016.

[37] Quanze Chen, Jonathan Bragg, Lydia B Chilton, and Dan S Weld. Cicero: Multi-turn, contextual argumentation for accurate crowdsourcing. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[38] Hao-Fei Cheng, Ruotong Wang, Zheng Zhang, Fiona O'Connell, Terrance Gray, F Maxwell Harper, and Haiyi Zhu. Explaining decision-making algorithms through

ui: Strategies to help non-expert stakeholders. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[39] Nalin Chhibber and Edith Law. Towards teachable conversational agents. *arXiv preprint arXiv:2102.10387*, 2021.

[40] Chun-Wei Chiang and Ming Yin. Exploring the effects of machine learning literacy interventions on laypeople's reliance on machine learning models. In *27th International Conference on Intelligent User Interfaces*, pages 148–161, 2022.

[41] Sue Clegg, Sally Bradley, and Karen Smith. 'i've had to swallow my pride': help seeking and self-esteem. *Higher Education Research & Development*, 25(2):101–113, 2006.

[42] Derrick Coetzee, Seongtaek Lim, Armando Fox, Bjorn Hartmann, and Marti A Hearst. Structuring interactions for large-scale synchronous peer learning. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 1139–1152, 2015.

[43] Nicola Cornally and Geraldine McCarthy. Help-seeking behaviour: A concept analysis. *International journal of nursing practice*, 17(3):280–288, 2011.

[44] Fabio Cuzzolin, Alice Morelli, Bogdan Cirstea, and Barbara J Sahakian. Knowing me, knowing you: theory of mind in ai. *Psychological medicine*, 50(7):1057–1061, 2020.

[45] Khanh-Hung Dang and Kim-Tuyen Cao. Towards reward-based spatial crowdsourcing. In *2013 International Conference on Control, Automation and Information Sciences (ICCAIS)*, pages 363–368. IEEE, 2013.

[46] Bella M DePaulo, WR Dull, James M Greenberg, and Gregory W Swaim. Are shy people reluctant to ask for help? *Journal of Personality and Social Psychology*, 56(5):834, 1989.

[47] Michael Desmond, Michael Muller, Zahra Ashktorab, Casey Dugan, Evelyn Duesterwald, Kristina Brimijoin, Catherine Finegan-Dollak, Michelle Brachman, Aabhas Sharma, Narendra Nath Joshi, et al. Increasing the speed and accuracy of data labeling through an ai assisted interface. In *26th International Conference on Intelligent User Interfaces*, pages 392–401, 2021.

[48] Stephan Dickert, Namika Sagara, and Paul Slovic. Affective motivations to help others: A two-stage model of donation decisions. *Journal of Behavioral Decision Making*, 24(4):361–376, 2011.

[49] Serkan Dincer and A Doğanay. The impact of pedagogical agent on learners' motivation and academic success. *Practice and Theory in Systems of Education*, 10(4):329–348, 2015.

[50] Mira Dontcheva, Robert R Morris, Joel R Brandt, and Elizabeth M Gerber. Combining crowdsourcing and learning to improve engagement and performance. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 3379–3388, 2014.

[51] Shayan Doroudi, Ece Kamar, and Emma Brunskill. Not everyone writes good examples but good examples can come from anywhere. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 12–21, 2019.

[52] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2623–2634, 2016.

[53] Ali Dorri, Salil S Kanhere, and Raja Jurdak. Multi-agent systems: A survey. *Ieee Access*, 6:28573–28593, 2018.

[54] John F Dovidio. Helping behavior and altruism: An empirical and conceptual overview. *Advances in experimental social psychology*, 17:361–427, 1984.

[55] Steven Dow, Anand Kulkarni, Brie Bunge, Truc Nguyen, Scott Klemmer, and Björn Hartmann. Shepherding the crowd: managing and providing feedback to crowd workers. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1669–1674. 2011.

[56] Steven Dow, Anand Kulkarni, Scott Klemmer, and Björn Hartmann. Shepherding the crowd yields better work. In *Proceedings of the ACM 2012 conference on computer supported cooperative work*, pages 1013–1022, 2012.

[57] John J Dudley and Per Ola Kristensson. A review of user interface design for interactive machine learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–37, 2018.

[58] Ahmed Elgammal, Yan Kang, and Milko Den Leeuw. Picasso, matisse, or a fake? automated analysis of drawings at the stroke level for attribution and authentication. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[59] Sean E Ellis and Dennis P Groth. A collaborative annotation system for data visualization. In *Proceedings of the working conference on Advanced Visual Interfaces*, pages 411–414, 2004.

[60] Eamonn Ferguson, Michael Taylor, David Keatley, Niall Flynn, and Claire Lawrence. Blood donors' helping behavior is driven by warm glow: More evidence for the blood donor benevolence hypothesis. *Transfusion*, 52(10):2189–2200, 2012.

[61] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4):681–694, 2020.

[62] Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of human-ai workflow on decision making in clinical imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 1362–1374, 2022.

[63] Karën Fort and Benoît Sagot. Influence of pre-annotation on pos-tagged corpus development. In *The fourth ACL linguistic annotation workshop*, pages 56–63, 2010.

[64] Ujwal Gadiraju and Stefan Dietze. Improving learning through achievement priming in crowdsourced information finding microtasks. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*, pages 105–114, 2017.

[65] Ujwal Gadiraju, Besnik Fetahu, and Ricardo Kawase. Training workers for improving performance in crowdsourcing microtasks. In *European Conference on Technology Enhanced Learning*, pages 100–114. Springer, 2015.

[66] Ya'akov Gal, Avi Segal, Ece Kamar, Eric Horvitz, Chris Lintott, and Mike Walmsley. A new workflow for human-ai collaboration in citizen science. In *Proceedings of the 2022 ACM Conference on Information Technology for Social Good*, pages 89–95, 2022.

[67] Amihai Glazer and Kai A Konrad. A signaling explanation for charity. *The American Economic Review*, 86(4):1019–1028, 1996.

[68] Ian Glover, Glenn Hardaker, and Zhijie Xu. Collaborative annotation system environment (case) for online learning. *Campus-Wide Information Systems*, 2004.

[69] Yuan Gong, Yu-An Chung, and James Glass. Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3292–3306, 2021.

[70] John E Grable and So-hyun Joo. A further examination of financial help-seeking behavior. *Journal of Financial Counseling and Planning*, 12(1):55, 2001.

[71] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 90–99, 2019.

[72] Ben Green and Yiling Chen. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–24, 2019.

[73] Michael D Greenberg, Matthew W Easterday, and Elizabeth M Gerber. Critiki: A scaffolded approach to gathering design feedback from paid crowdworkers. In *Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition*, pages 235–244, 2015.

[74] Amelia Gulliver, Kathleen M Griffiths, and Helen Christensen. Barriers and facilitators to mental health help-seeking for young elite athletes: a qualitative study. *BMC psychiatry*, 12(1):1–14, 2012.

[75] Feyza Merve Hafızoğlu and Sandip Sen. Comparing human trust attitudes towards human and agent teammates. In *Proceedings of the 8th International Conference on Human-Agent Interaction*, pages 50–59, 2020.

[76] Marcelo Guerra Hahn, Silvia Margarita Baldiris Navarro, Luis De La Fuente Valentín, and Daniel Burgos. A systematic review of the effects of automatic scoring and automatic feedback in educational settings. *IEEE Access*, 9:108190–108198, 2021.

[77] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. Forming effective human-ai teams: Building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948*, 2022.

[78] Patrick Hemmer, Max Schemmer, Michael Vössing, and Niklas Kühl. Human-ai complementarity in hybrid intelligence systems: A structured literature review. *PACIS 2021 Proceedings*, 2021.

[79] Chien-Ju Ho and Ming Yin. Working in pairs: Understanding the effects of worker interactions in crowdwork. *arXiv preprint arXiv:1810.09634*, 2018.

[80] Francisca AJ Hoogendijk. The oxyrhynchus papyri. *The Journal of Egyptian Archaeology*, 100(1):509–511, 2014.

[81] Eric Horvitz. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pages 159–166, 1999.

[82] Hsiao-Ying Huang, Michael Twidale, and Masooda Bashir. 'if you agree with me, do i trust you?': An examination of human-agent trust from a psychological perspective. In *Intelligent Systems and Applications: Proceedings of the 2019 Intelligent Systems Conference (IntelliSys) Volume 2*, pages 994–1013. Springer, 2020.

[83] Lizah Ismail. What net generation students really want: Determining library help-seeking preferences of undergraduates. *Reference services review*, 2010.

[84] Randolph L Jackson and Eileen Fagan. Collaboration and learning within immersive virtual reality. In *Proceedings of the third international conference on Collaborative virtual environments*, pages 83–92, 2000.

[85] C Centeio Jorge, Siddharth Mehrotra, ML Tielman, and CM Jonker. Trust should correspond to trustworthiness: a formalization of appropriate mutual trust in human-agent teams. In *22nd International Trust Workshop 2021*, 2021.

[86] Carolina Centeio Jorge, Myrthe L Tielman, and Catholijn M Jonker. Artificial trust as a tool in human-ai teams. In *2022 17th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pages 1155–1157. IEEE, 2022.

[87] Carolina Centeio Jorge, Myrthe L Tielman, and Catholijn M Jonker. Assessing artificial trust in human-agent teams: a conceptual model. In *Proceedings of the 22nd ACM International Conference on Intelligent Virtual Agents*, pages 1–3, 2022.

[88] Ece Kamar. Directions in hybrid intelligence: Complementing ai systems with human intelligence. In *IJCAI*, pages 4070–4073, 2016.

[89] Anna Kasunic, Chun-Wei Chiang, Geoff Kaufman, and Saiph Savage. Turker tales: Integrating tangential play into crowd work. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 21–34, 2019.

[90] Abhishek Kaushik, Vishal Bhat Ramachandra, and Gareth JF Jones. An interface for agent supported conversational search. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*, pages 452–456, 2020.

[91] Faizal Khan and Omar Reyad. Application of intelligent multi agent based systems for e-healthcare security. *arXiv preprint arXiv:2004.01256*, 2020.

[92] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F Welch. Does a digital assistant need a body? the influence of visual embodiment and social behavior on the perception of intelligent virtual agents in ar. In *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 105–114. IEEE, 2018.

[93] Yanghee Kim, Amy L Baylor, PALS Group, et al. Pedagogical agents as learning companions: The role of agent competency and type of interaction. *Educational technology research and development*, 54(3):223–243, 2006.

[94] Theodora Kontogianni, Michael Gygli, Jasper Uijlings, and Vittorio Ferrari. Continuous adaptation for interactive object segmentation by learning from corrections. In *European Conference on Computer Vision*, pages 579–596. Springer, 2020.

[95] Dominique Kost, Christian Fieseler, and Sut I Wong. Finding meaning in a hopeless place? the construction of meaningfulness in digital microwork. *Computers in Human Behavior*, 82:101–110, 2018.

[96] Adrian Krenzer, Kevin Makowski, Amar Hekalo, Daniel Fitting, Joel Troya, Wolfram G Zoller, Alexander Hann, and Frank Puppe. Fast machine learning annotation in the medical domain: a semi-automated video annotation tool for gastroenterologists. *BioMedical Engineering OnLine*, 21(1):1–23, 2022.

[97] Eva G Krumhuber, Dennis Küster, Shushi Namba, and Lina Skora. Human and machine validation of 14 databases of dynamic facial expressions. *Behavior research methods*, 53(2):686–701, 2021.

[98] Anand Kulkarni, Matthew Can, and Björn Hartmann. Collaboratively crowdsourcing workflows with turkomatic. In *Proceedings of the acm 2012 conference on computer supported cooperative work*, pages 1003–1012, 2012.

[99] Philipp Kulms and Stefan Kopp. The effect of embodiment and competence on trust and cooperation in human–agent interaction. In *International Conference on Intelligent Virtual Agents*, pages 75–84. Springer, 2016.

[100] Aparna A Labroo and Kelly Goldsmith. The dirty underbelly of prosocial behavior: Reconceptualizing greater good as an ecosystem with unintended consequences, 2021.

[101] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of human-ai decision making: a survey of empirical studies. *arXiv preprint arXiv:2112.11471*, 2021.

[102] Yi Lai, Atreyi Kankanhalli, and Desmond Ong. Human-ai collaboration in healthcare: A review and research agenda. 2021.

[103] Andrew S Lan, Divyanshu Vats, Andrew E Waters, and Richard G Baraniuk. Mathematical language processing: Automatic grading and feedback for open response mathematical questions. In *Proceedings of the second (2015) ACM conference on learning@ scale*, pages 167–176, 2015.

[104] Laura Larke, Sian Brooke, Huw Davies, Anoush Margaryan, and Vili Lehdonvirta. Skills formation and skills matching in online platform work. In *Proceedings of the British Sociological Association Annual Conference 2019: Abstracts by Session*, page 25. BSA Publications Ltd, 2019.

[105] Michiel Larmuseau, Michael Sluydts, Koenraad Theuwissen, Lode Duprez, Tom Dhaene, and Stefaan Cottenier. Race against the machine: can deep learning recognize microstructures as well as the trained human eye? *Scripta Materialia*, 193:33–37, 2021.

[106] Walter Lasecki and Jeffrey Bigham. Self-correcting crowds. In *CHI'12 Extended Abstracts on Human Factors in Computing Systems*, pages 2555–2560. 2012.

[107] Walter S Lasecki, Juho Kim, Nick Rafter, Onkur Sen, Jeffrey P Bigham, and Michael S Bernstein. Apparition: Crowdsourced user interfaces that come to life as you sketch them. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1925–1934, 2015.

[108] Yezdi Lashkari, Max Metral, and Pattie Maes. Collaborative interface agents. *Readings in agents*, pages 111–116, 1997.

[109] Annabel M Latham, Keeley A Crockett, David A McLean, Bruce Edmonds, and Karen O'shea. Oscar: An intelligent conversational agent tutor to estimate learning styles. In *International conference on fuzzy systems*, pages 1–8. IEEE, 2010.

[110] Daniella Laureiro-Martínez, Stefano Brusoni, and Maurizio Zollo. The neuroscientific foundations of the exploration- exploitation dilemma. *Journal of Neuroscience, Psychology, and Economics*, 3(2):95, 2010.

[111] Edith Law, Parastoo Baghaei Ravari, Nalin Chhibber, Dana Kulic, Stephanie Lin, Kevin D Pantasdo, Jessy Ceha, Sangho Suh, and Nicole Dillen. Curiosity notebook: A platform for learning by teaching conversational agents. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–9, 2020.

[112] Florian Laws, Christian Scheible, and Hinrich Schütze. Active learning with amazon mechanical turk. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1546–1556, 2011.

[113] Matthew Lease. On quality control and machine learning in crowdsourcing. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*. Citeseer, 2011.

[114] Marie-Eve LeBel, John Haverstock, Sayra Cristancho, Lucia van Eimeren, and Gavin Buckingham. Observational learning during simulation-based training in arthroscopy: is it useful to novices? *Journal of Surgical Education*, 75(1):222–230, 2018.

[115] Doris Jung-Lin Lee, Joanne Lo, Moonhyok Kim, and Eric Paulos. Crowdclass: Designing classification-based citizen science learning modules. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*, 2016.

[116] Fiona Lee. When the going gets tough, do the tough ask for help? help seeking and power motivation in organizations. *Organizational behavior and human decision processes*, 72(3):336–363, 1997.

[117] Min Hun Lee, Daniel P Siewiorek, Asim Smailagic, Alexandre Bernardino, and Sergi Bermúdez i Badia. Towards efficient annotations for a human-ai collaborative, clinical decision support system: A case study on physical stroke rehabilitation assessment. In *27th International Conference on Intelligent User Interfaces*, pages 4–14, 2022.

[118] Sang Won Lee, Rebecca Krosnick, Sun Young Park, Brandon Keelean, Sach Vaidya, Stephanie D O'Keefe, and Walter S Lasecki. Exploring real-time collaboration in crowd-powered systems through a ui design tool. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–23, 2018.

[119] David Lefevre and Benita Cox. Delayed instructional feedback may be more effective, but is this contrary to learners' preferences? *British Journal of Educational Technology*, 48(6):1357–1367, 2017.

[120] Vili Lehdonvirta, Anoush Margaryan, and HUW Davies. Skills formation and skills matching in online platform work: policies and practices for promoting crowdworkers' continuous learning (crowdlearn). 2019.

[121] Michael Lewis. Designing for human-agent interaction. *AI Magazine*, 19(2):67–67, 1998.

[122] Michael Lewis, Huao Li, and Katia Sycara. Deep learning, transparency, and trust in human robot teamwork. In *Trust in Human-Robot Interaction*, pages 321–352. Elsevier, 2021.

[123] Jingyi Li, Michelle X Zhou, Huahai Yang, and Gloria Mark. Confiding in and listening to virtual agents: The effect of personality. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, pages 275–286, 2017.

[124] YE Li and Nicholas Epley. When the best appears to be saved for last: Serial position effects on choice. *Journal of Behavioral Decision Making*, 22(4):378–389, 2009.

[125] Henry Lieberman. Autonomous interface agents. In *Proceedings of the ACM SIGCHI Conference on Human factors in computing systems*, pages 67–74, 1997.

[126] Huan Ling, Jun Gao, Amlan Kar, Wenzheng Chen, and Sanja Fidler. Fast interactive object annotation with curve-gcn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5257–5266, 2019.

[127] Dongyu Liu, Sarah Alnegheimish, Alexandra Zytek, and Kalyan Veeramachaneni. Mtv: Visual analytics for detecting, investigating, and annotating anomalies in multivariate time series. *arXiv preprint arXiv:2112.05734*, 2021.

[128] Han Liu, Vivian Lai, and Chenhao Tan. Understanding the effect of out-of-distribution examples and interactive explanations on human-ai decision making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW2):1–45, 2021.

[129] Patricia L Lockwood, Ayat Abdurahman, Anthony S Gabay, Daniel Drew, Marin Tamm, Masud Husain, and Matthew AJ Apps. Aging increases prosocial motivation for effort. *Psychological Science*, 32(5):668–681, 2021.

[130] Patricia L Lockwood, Mathilde Hamonet, Samuel H Zhang, Anya Ratnavel, Florentine U Salmony, Masud Husain, and Matthew AJ Apps. Prosocial apathy for helping others when effort is required. *Nature human behaviour*, 1(7):1–10, 2017.

[131] Kurt Luther, Amy Pavel, Wei Wu, Jari-lee Tolentino, Maneesh Agrawala, Björn Hartmann, and Steven P Dow. Crowdcrit: crowdsourcing and aggregating visual design critique. In *Proceedings of the companion publication of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 21–24, 2014.

[132] Lena Mamykina, Thomas N Smyth, Jill P Dimond, and Krzysztof Z Gajos. Learning from the crowd: Observational learning in crowdsourcing communities. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2635–2644, 2016.

[133] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, pages 879–893. PMLR, 2018.

[134] K.-K. Maninis, S. Caelles, J. Pont-Tuset, and L. Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018.

[135] Kevis-Kokitsi Maninis, Sergi Caelles, Jordi Pont-Tuset, and Luc Van Gool. Deep extreme cut: From extreme points to object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 616–625, 2018.

[136] Antonia Mantonakis, Pauline Rodero, Isabelle Lesschaeve, and Reid Hastie. Order in choice: Effects of serial position on preferences. *Psychological Science*, 20(11):1309–1312, 2009.

[137] Anoush Margaryan. Understanding crowdworkers' learning practices. 2016.

[138] Anoush Margaryan. Self-regulated learning in the crowd workplace. In *SIG 14: Learning and Professional Development*, 2018.

[139] Anoush Margaryan. Comparing crowdworkers' and conventional knowledge workers' self-regulated learning strategies in the workplace. *Human Computation*, 6:83–97, 2019.

[140] Anoush Margaryan. Comparing crowdworkers' and conventional knowledge workers' self-regulated learning strategies in the workplace. *Human Computation*, 6:83–97, 2019.

[141] Anoush Margaryan, Timothy Charlton, and Ujwal Gadiraju. Learning and skill development in online platform work. 2020.

[142] Ati Suci Dian Martha and Harry B Santoso. The design and impact of the pedagogical agent: A systematic literature review. *Journal of Educators Online*, 16(1):n1, 2019.

[143] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of management review*, 20(3):709–734, 1995.

[144] Robert R McCrae and Paul T Costa Jr. The five-factor theory of personality. 2008.

[145] Ree M Meertens and Rene Lion. Measuring an individual's tendency to take risks: the risk propensity scale 1. *Journal of applied social psychology*, 38(6):1506–1520, 2008.

[146] Fien Mertens, Esther de Groot, Loes Meijer, Johan Wens, Mary Gemma Cherry, Myriam Deveugele, Roger Damoiseaux, Ann Stes, and Peter Pype. Workplace learning through collaboration in primary healthcare: a beme realist review of what works, for whom and in what circumstances: Beme guide no. 46. *Medical teacher*, 40(2):117–134, 2018.

[147] Thaddeus Metz. Utilitarianism and the meaning of life. *Utilitas*, 15(1):50–70, 2003.

[148] Joseph E Michaelis and Bilge Mutlu. Reading socially: Transforming the in-home reading experience with a learning-companion robot. *Science Robotics*, 3(21), 2018.

[149] Stuart E Middleton. Interface agents: A review of the field. *arXiv preprint cs/0203012*, 2002.

[150] Nasrin Mohammadhasani, Hashem Fardanesh, Javad Hatami, Naser Mozayani, and Rosa Angela Fabio. The pedagogical agent enhances mathematics learning in adhd students. *Education and Information Technologies*, 23(6):2299–2308, 2018.

[151] Lillio Mok, Sasha Nanda, and Ashton Anderson. People perceive algorithmic assessments as less fair and trustworthy than identical human assessments. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–26, 2023.

[152] Kristen R Monroe, Michael C Barton, and Ute Klingemann. Altruism and the theory of rational action: Rescuers of jews in nazi europe. *Ethics*, 101(1):103–122, 1990.

[153] Maria Moundridou and Maria Virvou. Evaluating the persona effect of an interface agent in a tutoring system. *Journal of computer assisted learning*, 18(3):253–261, 2002.

[154] Seyed Mousa Golestaneh and Fahimeh Askari. Help-seeking or help avoidance: Important motivational, personality and metacognitive antecedents role in help-seeking and help-avoidance between normal and gifted students. *European Online Journal of Natural and Social Sciences: Proceedings*, 2(3 (s)):pp–3403, 2014.

[155] Yiftach Nagar and Thomas W Malone. Making business predictions by combining human and machine intelligence in prediction markets. Association for Information Systems, 2011.

[156] Gina Neff. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 2016.

[157] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, pages 340–350, 2021.

[158] David G Novick, Edith Elizalde, and Nathaniel Bean. Toward a more accurate view of when and how people seek help with computer applications. In *Proceedings of the 25th annual ACM international conference on Design of communication*, pages 95–102, 2007.

[159] Besmira Nushi, Ece Kamar, and Eric Horvitz. Towards accountable ai: Hybrid human-machine analyses for characterizing system failure. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 6, pages 126–135, 2018.

[160] James Onken, Reid Hastie, and William Revelle. Individual differences in the use of simplification strategies in a complex decision-making task. *Journal of Experimental Psychology: Human Perception and Performance*, 11(1):14, 1985.

[161] Dim P Papadopoulos, Jasper RR Uijlings, Frank Keller, and Vittorio Ferrari. Extreme clicking for efficient object annotation. In *Proceedings of the IEEE international conference on computer vision*, pages 4930–4939, 2017.

[162] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. A slow algorithm improves users' assessments of the algorithm's accuracy. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–15, 2019.

[163] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ digital medicine*, 2(1):111, 2019.

[164] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. *arXiv preprint arXiv:2202.11812*, 2022.

[165] Russell Perkins, Zahra Rezaei Khavas, and Paul Robinette. Trust calibration and trust respect: A method for building team cohesion in human robot teams. *arXiv preprint arXiv:2110.06809*, 2021.

[166] Isaac Pinyol and Jordi Sabater-Mir. Computational trust and reputation models for open multi-agent systems: a review. *Artificial Intelligence Review*, 40(1):1–25, 2013.

[167] Dennis L Poepsel and David A Schroeder. 13.5 helping and prosocial behavior. *Introduction to Psychology*, 2019.

[168] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, pages 1–52, 2021.

[169] Claudette Pretorius, Darragh McCashin, Naoise Kavanagh, and David Coyle. *Searching for Mental Health: A Mixed-Methods Study of Young People's Online Help-Seeking*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2020.

[170] Ishaani Priyadarshini and Chase Cotton. Ai cannot understand memes: Experiments with ocr and facial emotions. *CMC-COMPUTERS MATERIALS & CONTINUA*, 70(1):781–800, 2022.

[171] Lingyun Qiu and Izak Benbasat. Evaluating anthropomorphic product recommendation agents: A social relationship perspective to designing information systems. *Journal of management information systems*, 25(4):145–182, 2009.

[172] Neil Rabinowitz, Frank Perbet, Francis Song, Chiyuan Zhang, SM Ali Eslami, and Matthew Botvinick. Machine theory of mind. In *International conference on machine learning*, pages 4218–4227. PMLR, 2018.

[173] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction*, 35(5-6):413–451, 2020.

[174] Amy Rechkemmer and Ming Yin. Motivating novice crowd workers through goal setting: An investigation into the effects on complex crowdsourcing task training. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, pages 122–131, 2020.

[175] Amy Rechkemmer and Ming Yin. When confidence meets accuracy: Exploring the effects of multiple performance indicators on trust in machine learning models. In *CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2022.

[176] Ines Rehbein, Josef Ruppenhofer, and Caroline Sporleder. Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 19–26, 2009.

[177] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. Experimental evidence of effective human–ai collaboration in medical decision-making. *Scientific reports*, 12(1):14952, 2022.

[178] Debra Rickwood and Kerry Thomas. Conceptual measurement framework for help-seeking for mental health problems. *Psychology research and behavior management*, 5:173, 2012.

[179] Senjuti Basu Roy, Ioanna Lykourentzou, Saravanan Thirumuruganathan, Sihem Amer-Yahia, and Gautam Das. Task assignment optimization in knowledge-intensive crowdsourcing. *The VLDB Journal*, 24(4):467–491, 2015.

[180] J Philippe Rushton, Roland D Chrisjohn, and G Cynthia Fekken. The altruistic personality and the self-report altruism scale. *Personality and individual differences*, 2(4):293–302, 1981.

[181] Stephan Saalfeld, Albert Cardona, Volker Hartenstein, and Pavel Tomančák. Catmaid: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25(15):1984–1986, 2009.

[182] Laura R Saslow, Robb Willer, Matthew Feinberg, Paul K Piff, Katharine Clark, Dacher Keltner, and Sarina R Saturn. My brother's keeper? compassion predicts generosity more among less religious individuals. *Social Psychological and Personality Science*, 4(1):31–38, 2013.

[183] Claudia Sassenrath, Jacquie D Vorauer, and Sara D Hodges. The link between perspective-taking and prosociality—not as universal as you might think. *Current Opinion in Psychology*, 44:94–99, 2022.

[184] Mike Schaekermann. Human-ai interaction in the presence of ambiguity: From deliberation-based labeling to ambiguity-aware ai. 2020.

[185] Johannes Schiebener and Matthias Brand. Self-reported strategies in decisions under risk: role of feedback, reasoning abilities, executive functions, short-term-memory, and working memory. *Cognitive processing*, 16:401–416, 2015.

[186] Noah L Schroeder and Chad M Gotch. Persisting issues in pedagogical agent research. *Journal of Educational Computing Research*, 53(2):183–204, 2015.

[187] Silke Schworm and Hans Gruber. e-learning in universities: Supporting help-seeking processes by instructional prompts. *British Journal of Educational Technology*, 43(2):272–281, 2012.

[188] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 888–896, 2022.

[189] AmirAli Sharifi, Richard Zhao, and Duane Szafron. Learning companion behaviors using reinforcement learning in games. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 5, 2010.

[190] Robert Simpson, Kevin R Page, and David De Roure. Zooniverse: observing the world's largest citizen science platform. In *Proceedings of the 23rd international conference on world wide web*, pages 1049–1054, 2014.

[191] Peter Slattery, Richard Vidgen, and Patrick Finnegan. Winning heads and hearts? how websites encourage prosocial behaviour. *Behaviour & Information Technology*, 40(9):933–961, 2021.

[192] Richard L Solomon. The influence of work on behavior. *Psychological Bulletin*, 45(1):1, 1948.

[193] Francis Stevens and Katherine Taber. The neuroscience of empathy and compassion in pro-social behavior. *Neuropsychologia*, 159:107925, 2021.

[194] AA Stukas and EG Clary. Altruism and helping behavior. 2012.

[195] Ryo Suzuki, Niloufar Salehi, Michelle S Lam, Juan C Marroquin, and Michael S Bernstein. Atelier: Repurposing expert crowdsourcing tasks as micro-internships. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2645–2656, 2016.

[196] Nina Svenningsson and Montathar Faraon. Artificial intelligence in conversational agents: A study of factors related to perceived humanness in chatbots. In *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*, pages 151–161, 2019.

[197] Matthew I. Swindall, Gregory Croisdale, Chase C. Hunter, Ben Keener, Alex C. Williams, James H. Brusuelas, Nita Krevans, Melissa Sellew, Lucy Fortson, and John F. Wallin. Exploring learning approaches for ancient greekcharacter recognition with citizen science data. In *2021 17th International Conference on eScience (eScience)*, pages 128–137. IEEE.

[198] Matthew I Swindall, Gregory Croisdale, Chase C Hunter, Ben Keener, Alex C Williams, James H Brusuelas, Nita Krevans, Melissa Sellew, Lucy Fortson, and John F Wallin. Exploring learning approaches for ancient greek character recognition with citizen science data. In *2021 IEEE 17th International Conference on eScience (eScience)*, pages 128–137. IEEE, 2021.

[199] Wei Tang, Ming Yin, and Chien-Ju Ho. Leveraging peer communication to enhance crowdsourcing. In *The World Wide Web Conference*, pages 1794–1805, 2019.

[200] Athanasios Theofilatos, Cong Chen, and Constantinos Antoniou. Comparing machine learning and deep learning methods for real-time crash prediction. *Transportation research record*, 2673(8):169–178, 2019.

[201] Megan Tschannen-Moran and Wayne K Hoy. A multidisciplinary analysis of the nature, meaning, and measurement of trust. *Review of educational research*, 70(4):547–593, 2000.

[202] Jarvis Tse, Alex C. Williams, Joslin Goh, Timothy Player, James H. Brusuelas, and Edith Law. Mia: A mixed-initiative annotation platform. 2022.

[203] Jorge A Ramirez Uresti. Should i teach my computer peer? some issues in teaching a learning companion. In *International Conference on Intelligent Tutoring Systems*, pages 103–112. Springer, 2000.

[204] Wouter Van Atteveldt, Mariken ACG Van der Velden, and Mark Boukes. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140, 2021.

[205] Douwe van der Wal, Iny Jhun, Israa Laklouk, Jeff Nirschl, Lara Richer, Rebecca Rojansky, Talent Theparee, Joshua Wheeler, Jörg Sander, Felix Feng, et al. Biological data annotation via a human-augmenting ai-based labeling system. *NPJ digital medicine*, 4(1):145, 2021.

[206] Eva Vass and Karen Littleton. Peer collaboration and learning in the classroom. *International handbook of psychology in education*, pages 105–135, 2010.

[207] Jennifer Wortman Vaughan. Making better use of the crowd: How crowdsourcing can advance machine learning research. *J. Mach. Learn. Res.*, 18(1):7026–7071, 2017.

[208] B Vinoth Kumar, S Sabareeswaran, and G Madumitha. A decennary survey on artificial intelligence methods for image segmentation. In *Advanced Engineering Optimization Through Intelligent Techniques: Select Proceedings of AEOTIT 2018*, pages 291–311. Springer, 2020.

[209] Kailas Vodrahalli, Tobias Gerstenberg, and James Y Zou. Uncalibrated models can improve human-ai collaboration. *Advances in Neural Information Processing Systems*, 35:4004–4016, 2022.

[210] ME Walton, Steven W Kennerley, DM Bannerman, PEM Phillips, and Matthew FS Rushworth. Weighing up the benefits of work: behavioral and neural analyses of effort-related decision making. *Neural networks*, 19(8):1302–1314, 2006.

[211] Chih-Chien Wang and Chia-Hsin Wang. Helping others in online games: Prosocial behavior in cyberspace. *CyberPsychology & Behavior*, 11(3):344–346, 2008.

[212] Xinru Wang, Zhuoran Lu, and Ming Yin. Will you accept the ai recommendation? predicting human behavior in ai-assisted decision making. In *Proceedings of the ACM Web Conference 2022*, pages 1697–1708, 2022.

[213] M McLure Wasko and Samer Faraj. "it is what one does": why people participate and help others in electronic communities of practice. *The journal of strategic information systems*, 9(2-3):155–173, 2000.

[214] Alex C Williams, Hyrum D Carroll, John F Wallin, James Brusuelas, Lucy Fortson, Anne-Francoise Lamblin, and Haoyu Yu. Identification of ancient greek papyrus fragments using genetic sequence alignment algorithms. In *2014 IEEE 10th international conference on e-science*, volume 2, pages 5–10. IEEE, 2014.

[215] Alex C Williams, John F Wallin, Haoyu Yu, Marco Perale, Hyrum D Carroll, Anne-Francoise Lamblin, Lucy Fortson, Dirk Obbink, Chris J Lintott, and James H Brusuelas. A computational pipeline for crowdsourced transcriptions of ancient greek papyrus fragments. In *2014 IEEE International Conference on Big Data (Big Data)*, pages 100–105. IEEE, 2014.

[216] Rainer Winkler and Matthias Söllner. Unleashing the potential of chatbots in education: A state-of-the-art analysis. 2018.

[217] Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4, 2021.

[218] Qiong Wu, Chunyan Miao, and Zhiqi Shen. A curious learning companion in virtual learning environment. In *2012 IEEE International Conference on Fuzzy Systems*, pages 1–8. IEEE, 2012.

[219] Qiong Wu, Zhiqi Shen, and Chunyan Miao. Stimulating students' curiosity with a companion agent in virtual learning environments. In *EdMedia+ Innovate Learning*, pages 2401–2409. Association for the Advancement of Computing in Education (AACE), 2013.

[220] Daijin Yang, Yanpeng Zhou, Zhiyuan Zhang, Toby Jia-Jun Li, and Ray LC. Ai as an active writer: Interaction strategies with generated text in human-ai collaborative

fiction writing. In *Joint Proceedings of the ACM IUI Workshops*, volume 10. CEUR-WS Team, 2022.

[221] Runzhe Yang, Yexiang Xue, and Carla Gomes. Pedagogical value-aligned crowdsourcing: Inspiring the wisdom of crowds via interactive teaching. In *International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2018.

[222] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9657–9666, 2019.

[223] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–12, 2019.

[224] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Jianlong Zhou, and Fang Chen. Do i trust my machine teammate? an investigation from perception to decision. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 460–468, 2019.

[225] Zahra Zahedi and Subbarao Kambhampati. Human-ai symbiosis: A survey of current approaches. *arXiv preprint arXiv:2103.09990*, 2021.

[226] Zahra Zahedi, Sarath Sreedharan, and Subbarao Kambhampati. A mental-model centric landscape of human-ai symbiosis. *arXiv preprint arXiv:2202.09447*, 2022.

[227] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-ai teams and complementary expertise. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

[228] Haiyi Zhu, Steven P Dow, Robert E Kraut, and Aniket Kittur. Reviewing versus doing: Learning and performance in crowd assessment. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*, pages 1445–1455, 2014.