Addressing attentional issues in augmented reality with adaptive agents
Possibilities and challenges

Syiem, Brandon Victor; Kelly, Ryan M.; Dingler, Tilman; Goncalves, Jorge; Velloso, Eduardo

DOI
10.1016/j.ijhcs.2024.103324

Publication date
2024

Document Version
Final published version

Published in
International Journal of Human Computer Studies

Int. J. Human–Computer Studies 190 (2024) 103324

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Addressing attentional issues in augmented reality with adaptive agents: Possibilities and challenges

Brandon Victor Syiem [a,b,*], Ryan M. Kelly [c], Tilman Dingler [d], Jorge Goncalves [b], Eduardo Velloso [a]

[a] School of Computer Science, The University of Sydney, Camperdown, Sydney, 2006, New South Wales, Australia
[b] School of Computing and Information Systems, The University of Melbourne, Parkville, Melbourne, 3010, Victoria, Australia
[c] School of Computing Technologies, RMIT University, Melbourne, 3000, Victoria, Australia
[d] Department of Sustainable Design Engineering, Delft University of Technology, Mekelweg, Delft, 2628 CD, South Holland, Netherlands

ARTICLE INFO

Keywords:
Augmented reality
Attention
Adaptive agents
Artificial intelligence
Human–AI interaction
Sense-making

ABSTRACT

Recent work on augmented reality (AR) has explored the use of adaptive agents to overcome attentional issues that negatively impact task performance. However, despite positive technical evaluations, adaptive agents have shown no significant improvements to user task performance in AR. Furthermore, previous works have primarily evaluated such agents using abstract tasks. In this paper, we develop an agent that observes user behaviour and performs appropriate actions to mitigate attentional issues in a realistic sense-making task in AR. We employ mixed methods to evaluate our agent in a between-subject experiment (N=60) to understand the agent's effect on user task performance and behaviour. While we find no significant improvements in task performance, our analysis revealed that users' preferences and trust in the agent affected their receptiveness of the agent's recommendations. We discuss the pitfalls of autonomous agents and highlight the need to shift from designing better Human–AI interactions to better Human–AI collaborations.

1. Introduction

Augmented Reality (AR) offers a promising platform to support data analysis for sense-making (Subramonyam et al., 2019). However, the richness of AR environments brings additional complexities to the task and can hinder users' performance. For example, the additional virtual information presented in the 3D environment within an AR application can lead to *information overload* (Gebhardt et al., 2019). AR interactions have also been known to suffer from the *attentional tunnelling effect* (Wickens et al., 1993; Syiem et al., 2020), which causes users to excessively focus on task-related elements at the expense of other events and objects (Syiem et al., 2021). In addition, sense-making tasks — which involve searching for and filtering information, making decisions with incomplete or even wrong information, and generating hypotheses (Andrews et al., 2010) — also suffer from issues related to data overload and attention management (Pirolli and Card, 2005; Andrews et al., 2010). Therefore, while AR can offer rich interactions and data visualizations for sense-making, the combined attentional issues inherent to both the sense-making process and AR interactions can negatively affect task performance.

In this context, artificial intelligence (AI) agents offer a promising avenue for mitigating attentional issues and supporting users in optimizing their use of the AR environment. For example, AI agents can make preliminary inferences on the data, cluster similar data points, suggest data points for further inspection, or hide data irrelevant to the task at hand (Feit et al., 2020). Further, agents can monitor users' behaviours and use this information to provide personalized content and assistance through reinforcement learning algorithms (Gebhardt et al., 2019).

Effective collaboration with AI agents is important in complex decision-making tasks with unclear rules for success and failure. In such tasks, AI systems have been developed to assist humans rather than to complete the task on their own (Gebhardt et al., 2019; Koch et al., 2019; Feit et al., 2020). Understandably, prior works have used abstract tasks to evaluate and isolate the effect of such adaptive agents on AR task performance (Gebhardt et al., 2019; Lindlbauer et al., 2019). However, using abstract tasks, such as simple visual search tasks, removes the uncertainty and complexity in complex decision-making, where users may disagree with the agent. Moreover, despite positive technical evaluations, adaptive agents in previous work have led to no significant task performance improvement in interactive applications (Gebhardt et al., 2019; Lindlbauer et al., 2019; Feit et al., 2020). These reports are often accompanied by user comments on the increased complexity of the AI-integrated tool (Koch et al., 2019) or reluctance to accept the

* Corresponding author at: School of Computer Science, The University of Sydney, Camperdown, Sydney, 2006, New South Wales, Australia.
E-mail address: brandon.syiem@sydney.edu.au (B.V. Syiem).

https://doi.org/10.1016/j.ijhcs.2024.103324
Received 2 March 2023; Received in revised form 7 June 2024; Accepted 22 June 2024
Available online 1 July 2024

AI's assistance (Feit et al., 2020). Given these reports, we posit that a better understanding of the possibilities and challenges involved in the interaction between user and agent in a realistic sense-making task can help design interactions that harness AR's enhanced visualization and manipulation capabilities while minimizing the attentional costs incurred.

This paper aims to understand how adaptive agents can mitigate attentional issues in AR sense-making tasks and the challenges associated with interacting with the adaptive agent. We developed an agent that assists users by performing actions grounded in psychological theories of attention. The agent *clusters* related AR content to enable users to construct perceptual groups, aiding visual search (Treisman, 1982) and reducing attentional tunnelling (Wickens and Long, 1995) within groups. The agent further *highlights* or *collapses* content based on the content's inferred relevance to the task at hand, reducing perceptual load and freeing up attentional capacity for the user (Lavie, 1995; Cartwright-Finch and Lavie, 2007). The agent chooses which virtual content to adapt using a reinforcement learning approach based on the user's decision strategy and intention by observing their gaze data (Feit et al., 2020; Gebhardt et al., 2019) and direct manipulations of the content. The agent receives positive feedback if the user finds and interacts with relevant virtual content that the agent has highlighted or if the user ignores content that the agent has collapsed. The agent receives negative feedback if the user interacts with content that the agent has collapsed or if the user ignores content that the agent has highlighted.

We conducted a between-participants experiment (N = 60) to explore users' behaviours, perceptions, and performance in a sense-making task using three experimental conditions: an analogue baseline with paper-based post-it notes (Paper), an interactive HoloLens v2 application in which users could interact with virtual post-it notes (Unassisted AR), and an AI-assisted version of the same application that incorporates our agent (Assisted AR). Our experiment was designed to isolate the effects of AR and AI compared to the baseline, focusing on how users collaborated with the agent to accomplish the task.

Our results suggest that despite evidence for the high accuracy of the agent's recommendations, the agent's assistance did not significantly improve task performance compared to the unassisted condition. It did, however, lead to a larger variation in performance, suggesting that the participants who were more receptive to the agent's recommendations outperformed those who were reluctant to consider the agent's assistance. Further analysis revealed three key differences that affected participants' acceptance of the agent's recommendations: (i) user *preferences* about the adaptations offered by the agent, (ii) *trust* in the agent's actions, and (iii) *knowledge* of the agent's capabilities. Our work contributes insights into the design of adaptive agents, highlights the challenges and potential of applying adaptive solutions for reducing attentional issues in sense-making tasks, and demonstrates the need for interaction designs where humans not only react to the actions of an AI but jointly collaborate with it to accomplish tasks.

2. Related work

2.1. Attentional issues in augmented reality

The increasing adoption of AR technology in commercial applications has brought to light a range of attentional issues reportedly caused by AR (Wagner-Greene et al., 2017; Ayers et al., 2016; Gebhardt et al., 2019; Syiem et al., 2020). These issues often cause users to miss events or objects (Wagner-Greene et al., 2017; Ayers et al., 2016; Syiem et al., 2021), hinder task performance (Gebhardt et al., 2019), and disrupt the user experience (Syiem et al., 2020). Among these issues, *information overload* and *attentional tunnelling* are the most well-understood issues in the context of AR (Lindlbauer et al., 2019; Gebhardt et al., 2019; Feit et al., 2020; Syiem et al., 2021).

Though AR presents the opportunity to enhance and add contextual information needed to perform tasks, the additional virtual information can also overwhelm the user and hinder their efforts rather than support them (Gebhardt et al., 2019). Such information overload can occur when the user is presented with more information than they can process at any given time (Schick et al., 1990). Previous works have therefore attempted only to show virtual content immediately relevant to the task at hand (Lindlbauer et al., 2019; Gebhardt et al., 2019), since the presence of irrelevant content has been shown to reduce user task performance (Tatzgern et al., 2016).

Task-related content in AR applications has also been known to cause the attentional tunnelling effect (Syiem et al., 2021). This effect occurs when users excessively focus their attention on a specific channel of information, hypothesis, or task goal, which can result in neglecting other channels of information, hypotheses, or tasks (Wickens and Alexander, 2009). Like information overload, attentional tunnelling has been reported to reduce users' performance in tasks that involve visual search (Kortschot and Jamieson, 2019; Syiem et al., 2021).

Reducing the perceptual load of task-related content in AR applications can reduce attentional tunnelling and the related phenomenon of *inattentional blindness* (Cartwright-Finch and Lavie, 2007). Macdonald and Lavie (2011) define perceptual load as "the amount of information involved in the perceptual processing of task stimuli" and operationalize it in terms of the amount of task-related content or the perceptual requirements of the task in the same content (simple versus complex discrimination task) (Macdonald and Lavie, 2011, p.1780). This suggests that we can mitigate the adverse effects of attentional tunnelling in AR by reducing the amount of task-related content and easing the visual search for task-related content. An additional measure that can be taken to reduce attentional tunnelling and aid in visual search in AR is to enable perceptual grouping of virtual content that needs to be attended to efficiently (Treisman, 1982; Wickens and Long, 1995). This is because the effects of attentional tunnelling can be reduced when the attended item and the unattended item(s) belong to the same perceptual group, as opposed to when the users perceptually group them as separate objects (Wickens and Long, 1995).

In this paper, we develop an adaptive AI agent to mitigate information overload and attentional tunnelling in AR by intelligently reducing the user's perceptual load. The agent does this by reducing the amount of virtual content concurrently presented to the user and easing the visual scanning task required to find relevant information. The agent also attempts to enable users to perceptually group virtual content with similar information to reduce attentional tunnelling.

2.2. Attentional issues in sense-making tasks

Pirolli and Card describe sense-making as an iterative process that involves two distinct loops, namely the *foraging* loop, which broadly involves searching for information and filtering relevant from irrelevant information, and the *sense-making* loop, which involves building a mental model (or generating hypotheses) from the gathered information (Pirolli and Card, 2005). They further identify issues related to data overload and attention management in the sense-making process (Pirolli and Card, 2005; Andrews et al., 2010). Prior works have attempted to mitigate the attentional issues involved in sense-making by exploring technological solutions, such as large, high-resolution displays (Andrews et al., 2010), the timing of clue interruptions based on user's state of arousal (Goyal and Fussell, 2017) and novel interfaces for interacting with the sense-making tasks (Goyal and Fussell, 2016).

In addition, the individual processes involved in a sense-making task have been associated with various attentional issues (Pirolli and Card, 2005). Table 1 shows examples of previous studies that demonstrate how the sub-tasks involved in a sense-making process induce attentional issues. As we focus on sense-making in AR environments, Table 1 only summarizes previous studies involving tasks with visual information sources.

Table 1
Table showing previous works that employ the sub-tasks of searching, filtering and/or generating hypothesis, involved in the sense-making process to explore or induce attentional phenomena. The attentional phenomena discussed in these works are related to information overload, perceptual load or cognitive load. The table gives a brief description of the task employed and how the task relates to an attentional phenomenon. The table further details how the task is associated with the sub-tasks involved in the sense-making process as described by Pirolli and Card (2005).

	Task description	Foraging loop		Sense-making loop
		Search task	Filter task	Generate hypothesis
Goyal and Fussell (2016)	Employed a collaborative crime-solving task where pairs of participants were tasked with finding a serial killer from 3 cold murder case documents. The task was used to explore a novel system that aimed to reduce *attentional tunnelling* by enabling participants to explicitly track their collaborator's hypothesis. ***Search Task:*** Find information about the crime from multiple given clues. ***Filter Task:*** Filter out false testimonies, irrelevant information, etc. from relevant clues that share the same visual features (digital sticky notes). ***Generate Hypothesis:*** Build hypothesis around the murderer, time of murder, weapon used, etc.	✓	✓	✓
Goyal and Fussell (2017)	Employed a crime solving task to explore the effects of timing of clue interruptions, where participant's receive a clue at specific times based on their state of arousal. Timing of receiving a clue was hypothesized to affect *attentional tunnelling* into one's own mental model or attentional tunnelling into the newly acquired clue. ***Search Task:*** Find information about the crime from multiple given clues. ***Filter Task:*** Filter out false testimonies, irrelevant information, etc. from relevant clues presented in the same textual format. ***Generate Hypothesis:*** Build hypothesis around the murderer, time of murder, weapon used, etc.	✓	✓	✓
Lindlbauer et al. (2019)	Employed a search task for a target icon amongst 30 distractor icons which induced low *cognitive load*. ***Search Task:*** Search for a given target icon. ***Filter Task:*** Filter out non-target icons.	✓	✓	✗
Gebhardt et al. (2019)	Employed a visual search task to evaluate their adaptive system that aimed to reduce *information overload* by reducing the amount of content presented in AR. ***Search Task:*** Find objects with specific label requirements (matching string or maximum/minimum number). ***Filter Task:*** Filter out objects with irrelevant labels. Objects are easier or harder to distinguish based on whether they possess attentive or pre-attentive features.	✓	✓	✗
Cartwright-Finch and Lavie (2007)	Conducted experiments exploring *perceptual load* using a visual search task efficiency for finding target amongst 5 distractors. Found that *perceptual load* is higher when distractors share similar features to the target. ***Search Task:*** Search for target amongst 5 distractors. Search efficiency was compared between finding target between distractors with similar features and placeholder distractors. ***Filter Task:*** Filter out distractors.	✓	✓	✗
Greene et al. (2017)	Conducted an experiment to see how *perceptual load*, induced by *visual clutter*, affects detection of peripheral objects in a video screening task. Found that video depicting a more cluttered environment (51 irrelevant objects as opposed to 13 irrelevant objects) induced higher perceptual load and reduced detection of peripheral character. ***Filter Task:*** Filter out relevant from irrelevant content from a video depicting a bank robbery.	✗	✓	✗
Kortschot and Jamieson (2019)	Conducted an experiment consisting of a search task where subsequent targets would appear individually once the previous target was found. All targets, except the last, would appear on a pre-determined sub-section (unknown to participants) of the canvas to induce *attentional tunnelling*. Search time for the last target was compared to a baseline condition where all targets appeared randomly on the canvas. ***Search Task:*** Find and accurately 'tag' target as 'horizontal', 'vertical' or 'diagonal' target displayed on the digital canvas. ***Generate Hypothesis:*** A hypothesis of where the next target would appear on the canvas was induced in the experimental condition.	✓	✗	✓

To ensure that the sense-making task used in our experiment involves a degree of attentional challenge comparable to the works discussed in Table 1, we employ a crime-solving task similar to the task used by Goyal and Fussell (Goyal and Fussell, 2016, 2017). A crime-solving task involves all sub-tasks present in a sense-making process, i.e., searching for clues, filtering out relevant from irrelevant clues, and

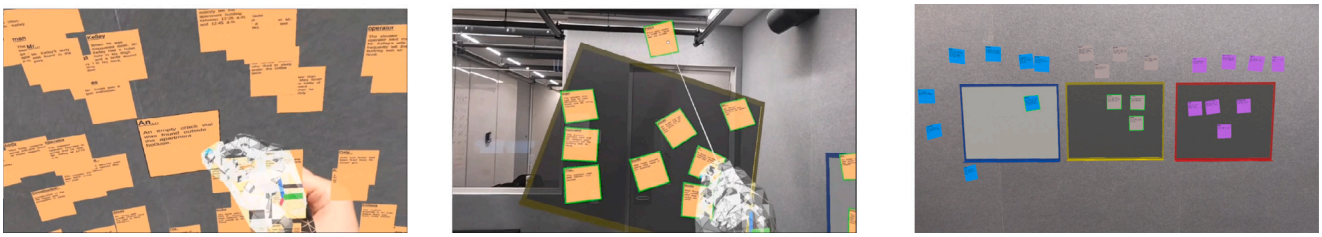

Fig. 1. Images showing (a) a participant using the near interaction to interact with virtual post-it notes, (b) a participant using the far interaction to interact with virtual post-it notes and (c) Notes similar to attached notes on boards being clustered with the same colour.

generating hypotheses on possible suspects and details about the crime (see Table 1). To ensure that our task is attentionally demanding, we used several distinct clues that exceed the number used in the works presented in Table 1. Specifically, we ensured that the number of clues presented in our task (60) exceeds the number used in previous works exploring information overload and perceptual load (Greene et al., 2017) (51 clues) and cognitive load (Lindlbauer et al., 2019) (30). All clues are also presented using the same visual features to increase the perceptual load on the user (Cartwright-Finch and Lavie, 2007) (see Table 1 for details).

2.3. AI adaptive methods for assisting user tasks

Several adaptive AI agents have been developed to *assist* users in decision-making tasks (Feit et al., 2020; Koch et al., 2019) and in tasks presented in Mixed Reality (MR) (Gebhardt et al., 2019; Lindlbauer et al., 2019). These adaptive agents implicitly observe user data, such as gaze (Gebhardt et al., 2019; Feit et al., 2020) or cognitive load (Lindlbauer et al., 2019), and adapt the user interface to support the user's task.

For example, Lindlbauer et al. (2019) created a system to automatically adapt the user interface in an MR application where the user has multiple windows open for different tasks. The system automatically adjusts the visibility and information presented on the windows based on the user's cognitive load and current task. However, the researchers did not find any improvement in task performance using this adaptive method in a simple counting and visual search task.

Similarly, Gebhardt et al. (2019) designed an adaptive agent to mitigate information overload by reducing visual clutter in an MR application. They employed a model-free reinforcement learning agent to determine user intentions based on gaze data. The agent then minimized the amount of virtual content presented to the user based on inferred intentions. They also found that the agent did not significantly improve users' performance in a visual search task.

To explore how adaptive methods can be used to assist users in a realistic decision-making task, Feit et al. (2020) explored how users' gaze data could detect information relevance in an apartment sharing website. They argued that gaze behaviour varies greatly between individuals and that ground truth data can be challenging to obtain in decision-making tasks. As such, they identified six eye-tracking metrics to predict the relevance of interface elements for individual users. Despite their users reporting high accuracy for their proposed method, they did not observe a significant improvement in the participants' decision-making task performance. However, their findings cannot be directly applied to a realistic sense-making task in AR due to differences in dimensionality and the added attentional issues observed in AR.

The lack of performance improvement when using these adaptive methods despite their positive technical evaluations (Lindlbauer et al., 2019; Gebhardt et al., 2019; Feit et al., 2020) hints at a different problem altogether: challenges in Human–AI interactions. The challenges associated with effectively interacting with AI agents have been widely acknowledged (Amershi et al., 2019; Yang et al., 2020). For example, unpredictable AI actions and behaviour can confuse users and lead users to abandon AI technology (De Graaf et al., 2017). Similarly, automated filtering of content by AI can lead to unwanted information hiding (Amershi et al., 2019). To address these issues, recent works have proposed guidelines to support users in understanding AI behaviour and enabling users to undo AI actions (Amershi et al., 2019). However, these problems persist in recent implementations of adaptive agents (Yang et al., 2020). Our work aims to understand these challenges and possible opportunities in interacting with AI agents designed to mitigate attentional issues in AR decision-making tasks.

3. System design and implementation

To explore how adaptive agents can mitigate attentional issues in AR sense-making tasks and the challenges associated with interacting with the agent, we developed (1) an AR application to present sense-making information as interactive AR content, and (2) an adaptive agent designed to mitigate the attentional issues involved in AR interactions and the sense-making process. The agent observes users' gaze data and the current task state to determine the *immediate* relevance of different pieces of virtual content involved in the task. The agent then adapts the virtual content to reduce perceptual load and reduces visual clutter to minimize attentional tunnelling and information overload. Because both of the latter effects reduce user performance in visual search tasks (Kortschot and Jamieson, 2019; Tatzgern et al., 2016; Gebhardt et al., 2019), we expect users to find relevant task information faster with the help of the adaptive agent. Additionally, we provide users with interactions to undo the agent's actions for adaptations that hide information from the user to adhere with guidelines presented in prior work on Human–AI interaction (Amershi et al., 2019).

The agent runs on a custom-built decision-making support application. The application renders text snippets obtained from a CSV file as virtual post-it notes. The user can then move these notes around the virtual space and cluster related notes on virtual boards (see Fig. 1). We chose a sense-making activity as our decision-making task because it requires substantial human interpretation (making it difficult to fully automate) while offering enough complexity for the assistance of an artificial agent to bring meaningful advantages.

3.1. AR application

We developed the AR application in Unity for the Microsoft HoloLens v2, using Microsoft's Mixed Reality Toolkit version 2 (MRTKv2). Once started, the application loads the text snippets and displays them as virtual post-it notes randomly arranged around the virtual environment. In addition to the post-it notes, the user can create virtual *boards* using speech commands. Users can pick up post-its with a pinch gesture, move/rotate them around the space, and drop them onto these boards (see Fig. 1(a) and (b)). If a post-it is attached to a board, and the board is manipulated, the board and all attached post-its move together in 3D space.

The AR application supports two kinds of virtual boards: *relevant cluster* boards and *garbage* boards. Relevant cluster boards (created with the voice command "create board") are intended for grouping related notes relevant to the user. Garbage boards (created with the voice command "garbage board") are meant to store notes the user deems

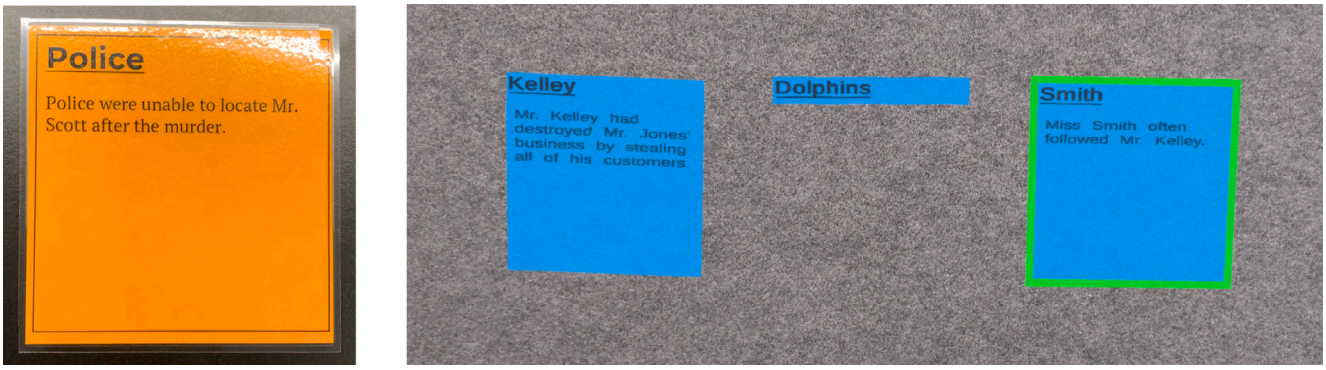

Fig. 2. Images of (a) a physical post-it note (b) virtual post-it notes in (i) Normal State, (ii) Folded State, and (iii) Outlined State.

irrelevant. Both kinds of boards can be destroyed by saying "destroy". When a board is destroyed, the post-it notes attached to it are *not* destroyed and are left with their positions unchanged. The "destroy" command enables users to remove digital clutter in the event that an unnecessary amount of virtual boards were created.

The application also supports two 3D interaction techniques for manipulating notes and boards. The user can directly manipulate them by picking them up and dropping them off (*near interaction*). The user can also interact with notes and boards from a distance by pointing at them — which causes a virtual ray to extend from their finger in the direction they are pointing — and pinching (*far interaction*). All object transformations are isometric.

Our application collects user behaviour and task data that inform the agent's behaviour. It sends user intention and task-state data in fixed time windows (which we call "communication window") to the adaptive agent. A 5-second window was chosen based on pilot testing (see Appendix C). This window allowed enough time for users to meaningfully interact with the application's content (find and view relevant notes, and manipulate notes). Additionally, the 5-second window avoids prompting the agent for excessive adaptations that could visually distract users, while maintaining a short enough duration that the user's immediate task, operationalized through gaze and interaction data, does not change within the set window time. We approximate user intentions based on their eye-tracking data (Gebhardt et al., 2019; Feit et al., 2020), which we record using the in-built eye-tracker in the HoloLens 2. We collect task-state data corresponding to the mapping between post-it notes and virtual boards. Specifically, the application sends the following data to the agent:

- Gaze dwell time for each post-it note. This is the ratio between the time spent looking at a post-it note and the communication window.
- Saccadic ins for each post-it note. This is the ratio between the number of times the users gaze has *entered* a post-it note and the total number of times the users gaze has entered any post-it note within a communication window.
- Post-it note to virtual board map. This is a map indicating which post-it note is attached to which virtual board (if any).

The application receives actions from the agent which dictate how to adapt the visualization of the post-it notes based on current user intention data and the application task-state data.

3.2. Adaptive agent

The adaptive agent was written in Python 3 and ran on a dedicated machine separate from the AR application. The agent has four primary functions: (i) it uses Natural Language Processing (NLP) techniques to generate headings for the post-it notes (see Fig. 2 for examples); (ii) it converts the content of the notes into sentence embeddings (a technique in which sentences are numerically represented as a vector of real numbers), and uses this data along with the mapping between post-its and boards to cluster post-it notes based on similarity; (iii) it feeds data sent from the AR applications along with the sentence embeddings to a reinforcement learning (RL) agent that outputs a set of actions to adapt each post-it note; and (iv) it communicates with the AR application to receive data, to send actions back to the application and to cluster information.

Natural language processing module. The NLP module of the adaptive agent primarily handles two main functions:

1. It generates a header for each clue on a post-it note. Specifically, the Python library spaCy[1] is used to tokenize each clue and find a list of subjects within the tokens. The header is set to the first detected subject within the clue. For example, the header "Smith" was generated for the clue "Miss Smith often followed Mr Kelley". If no subject is detected, the header is set to the first word in the sentence, followed by a trailing ellipsis (...). These headers were used to visualize an adapted state of a post-it note, and not used as input to the RL agent.
2. It creates a 768-dimension vector representation of each clue using Google's BERT transformer-based language model (Reimers and Gurevych, 2019). These vectors, called sentence embeddings, are used to measure similarities between different notes using angular similarity (Cer et al., 2018):

$$ang\_sim(u, v) = 1 - arccos\left(\frac{u.v}{\|u\|.\|v\|}\right) \quad (1)$$

where u and v are the vector representation of two clues. Sentence embeddings were also used as input for the RL agent.

Clustering module. As the user attaches notes to the virtual boards, the agent suggests possible clusters to which other unattached notes might belong. Formally, the agent creates k clusters, where k is the number of virtual boards (relevant cluster and garbage boards) with post-its attached to them.

It then calculates the mean for each cluster i (μ_i) over the sentence embeddings of the clues attached to the virtual board with ID i (where $i \epsilon [1, k]$). Each unattached post-it note is then grouped into a cluster based on the angular similarity between the notes embedding and the mean of the different clusters, i.e., the cluster C_i to which an unattached note with embedding n is assigned is given by:

$$C_i = \underset{i}{argmax}(ang\_sim(n, \mu_i)) \qquad \forall i \in 1..k \quad (2)$$

where ang_sim() is defined in Eq. (1).

[1] spaCy is an open-source library for advanced Natural Language Processing (NLP) in Python — https://spacy.io/

The clustering module outputs a list of all post-it note IDs with their assigned cluster number, i.e., a list of mappings between notes and boards. Clustering is then visualized in the AR application using distinct colours for individual clusters (see Fig. 1(c)). We chose this approach because the use of colours enables effective perceptual grouping (Baylis and Driver, 1992; Fox, 1998), which in turn mitigates attentional tunnelling within perceptual groups (Wickens and Long, 1995).

Reinforcement learning module. In addition to suggesting cluster membership, the agent tries to focus users' attention on the task at hand in two ways. First, it tries to minimize the visual clutter in the virtual environment by collapsing notes into a smaller version containing only their headings, if they are deemed irrelevant to the current analysis. Second, it tries to draw attention to notes that are particularly relevant by highlighting them with a distinct outline (see Fig. 2(b)).

The agent decides on one of three different actions for each virtual post-it note based on the user's gaze data, the sentence embeddings of clues, and the current task state (user-selected relevant notes). The agent either (i) collapses a note to reduce information overload and perceptual load (by reducing the number of task-related content (Macdonald and Lavie, 2011)) if the agent determines it to be irrelevant to the user — Folded State; (ii) outlines a note using highlights (Feit et al., 2020), to reduce perceptual load by visually guiding (Biggs et al., 2015) and easing the search for relevant post-it notes if it determines the note to be relevant to the user – Outlined state; or (iii) displays all content of the note without outlining or collapsing the note if the content is neither highly relevant nor irrelevant to the user — Normal State (see Fig. 2(b)). The relevance of specific notes is highly dependent on individual user strategies and intentions and can be challenging to model. As such, we use a model-free reinforcement learning (RL) agent to predict the relevance of different notes based on users' gaze data and notes that the user has deemed relevant during the task (by placing these notes on the relevant cluster boards).

To inform the RL algorithm, we first define an observation space — what the agent needs to keep track of — and an action space — the set of possible actions that the RL agent can perform. In our application, the observation space consists of gaze data for each virtual post-it note — the time spent looking at the notes and the number of times the user's gaze entered a note (i.e. gaze dwell time and saccadic ins; 2 parameters) — the sentence embeddings for each note (768 parameters), the current action-based state of the note — Folded, Outlined or Normal state (1 parameter) — and the note's relevance as indicated by the user (by attaching the note on any relevant cluster board) — relevant or irrelevant (1 parameter).

This produces an observation state space of 772 parameters (768 + 2 + 1 + 1) for each note. In the study task, we had 60 notes, resulting in 60 x 772 parameters. The action state is one of three actions for each of the 60 post-it notes, i.e., 3^{60} different possible action states. We use a deep neural network based RL agent capable of handling such extensive observation and action spaces called Deep Deterministic Policy Gradient (DDPG) (Casas, 2017).

Our DDPG agent was written using `tensorflow 2` (Abadi et al., 2015) through the `keras` API (Chollet et al., 2015). DDPG consists of two neural networks: an actor network that takes in the observation state and outputs an action state and a critic network that predicts the quality (good or bad) of the generated action given the observation state (Chollet et al., 2015). Our implementation of the actor network uses a softmax function that outputs a probability distribution across the 3 actions for each of the 60 post-it notes (i.e., a 3 x 60 action state). The action with the highest probability for each post-it is then selected and sent to our AR application. The critic network uses a set of defined rewards to determine the quality of an action over a given observation state.

In our implementation, a reward is given on the following observation state received from the AR application after the agent has sent an action set to the AR application. This enables the agent to observe how the user responds to the agent's actions and adjust its decision-making accordingly. A positive reward is given when an outlined post-it note is observed to be relevant or when a folded post-it note is observed to be irrelevant. A negative reward is given when an outlined note is considered irrelevant, or a folded note is considered relevant. Notes in the normal state receive a small positive reward regardless of whether they are observed to be relevant or irrelevant. The small positive reward for notes in the normal state discourages the agent from frequently switching states between folded and outlined. This was implemented after observing that the flickering of notes between the outlined and folded states hindered the usage of the system during a pilot study. For an action a taken by the agent to change state s to state s', the reward received by the agent when it observes the next state s'' after the user has seen and interacted with state s', can be expressed as:

$$r(s,a,s',s'') = \begin{cases} +r_o & if\ outlined\ note\ is\ relevant\ in\ s'' \\ -r_o & if\ outlined\ note\ is\ irrelevant\ in\ s'' \\ +r_n & if\ normal\ note\ is\ relevant\ or\ irrelevant\ in\ s'' \\ +r_{fi} & if\ folded\ note\ is\ irrelevant\ in\ s'' \\ -r_{fr} & if\ folded\ note\ is\ relevant\ in\ s'' \end{cases} \quad (3)$$

Where, $r_o > r_{fr} \gg r_{fi} \gg r_n$. These reward values are relative and based on the limited number of post-it notes a user can interact with within the 5-second window (where the agent observes the system's state). Specifically, the positive reward for an observed relevant outlined post-it note ($+r_o$) is equal to the negative reward for an observed irrelevant outlined post-it note ($-r_o$). This is so that the agent does not outline too many notes and attempts to optimize the number of notes a user can select within the 5-second window (as it will receive a large negative reward). Similarly, when the user selects a folded note, the agent gets a slightly smaller negative reward ($-r_{fr}$). The negative reward is smaller than r_o to encourage the agent to fold more than the number of notes it should outline. Both r_o and r_{fr} are much larger than the reward for observing an irrelevant folded note $+r_{fi}$ because there are more notes that the user will not interact with than will interact with. A smaller r_{fi} stops the agent from folding all notes to receive large rewards. Finally, a very small positive reward (r_n) is awarded to the agent for leaving the notes in the normal state. This encourages the agent not to flick between outlined and folded states, especially when the agent is unsure how relevant the note is to the user.

Finally, to stop our agent from performing completely random actions during the initial phases of the experiment, we trained the agent using a sense-making data-set[2] separate from the one used for our experiment on a naive simulated user. The simulated user first selects [0,m] random post-it notes as relevant, where 'm' is the maximum number of notes that can be selected within the 5-second window and is based on a previous pilot test. The agent then performs actions based on the observed state of the system. The simulated user now has to decide which notes to mark as relevant for the next window. It does this by first creating two clusters; a 'relevant' cluster with the centroid equal to the mean of the sentence embeddings of notes that have been marked as relevant and an 'irrelevant' cluster with the centroid equal to the mean of the sentence embeddings of notes that have been *not* marked as relevant. The simulated user then selects [0,k] notes from the irrelevant cluster to mark as relevant based on the angular similarity (Eq. (1)) between the notes in the irrelevant cluster to the centroid of the relevant cluster. There is also a small probability (1/5) that users pick 'l' notes randomly (not based on similarity) to place on the relevant cluster board (where $k+l = m$). This simulates how users can switch between different topics within the sense-making task. The updated state is then used to determine the reward for the agent as per Eq. (3). Note that the eye tracking data (Gaze dwell time and Saccadic

[2] https://peterpappas.com/images/2010/08/Bank-Robbery-1.pdf

Table 2
Summary Statistics for normalized rewards the random solver and adaptive agent.

System	N	Mean	SD
Random solver	20	0.119	0.071
Adaptive agent	20	0.929	0.054

Ins) are spread uniformly across all post-it notes, i.e., this data is not used to *initially* train the system.

We stress that the described training step is only used to stop the agent from making completely random actions during the initial phases of the study. The naively trained state of the agent is exposed to further training when a user uses the system (which exposes it to eye-tracking data and more personalized means of selecting relevant notes). This allows the agent to adapt to each user's *personal* strategies of solving the sense-making task.[3]

4. Technical evaluation

To assess whether the agent optimizes for the reward function presented in Section 3.2, we compared the performance of our adaptive agent to a baseline random solver. The random solver performs actions randomly, and was chosen as the baseline in the absence of other systems designed to assist users with realistic sense-making tasks in AR. We gathered and accumulated the rewards using our agent and the random solver for 20 episodic tests each. Each episode is equivalent to 30 min of task time, which is based on reported completion times for the chosen task in prior work (Wozniak et al., 2016). The tests were run using the simulated user on the actual experimental sense-making dataset by Stanford and Stanford (Stanford and Stanford, 1969), as opposed to the training dataset described in Section 3.2. The use of a dataset other than the one used to train the agent ensures that the rewards accrued by the agent is not a result of over-fitting and allows for a fair comparison against the random solver.

Table 2 shows the summary statistics for the normalized rewards[4] of the adaptive agent and random solver. A Mann–Whitney's U test showed the reward accumulated by the adaptive agent was significantly larger than the reward accumulated by the random solver ($U = 400, \quad Z = 5.41, \quad p = 1.451e-11$). The size of the effect was large, $r = 0.85$. The result shows that the adaptive agent achieves a significantly larger reward than the random solver even on the experimental dataset that was not used in training.

To further show that our agent learns to optimize rewards over time, we collected the rewards achieved by the agent per training episode and plotted the normalized rewards against increasing number of episodes trained. Fig. 3 shows the plot between the normalized rewards achieved in a single episode over the number of episodes trained on. An increasing trend in the reward indicates that the agent learns to better optimize for our reward function as it trains on more episodes. The troughs can be explained by the probabilistic nature of the simulated user to perform a random action (described in Section 3.2).

5. Method

We conducted an experiment to understand how adaptive agents can be used to mitigate attentional issues in AR sense-making tasks and the challenges associated with interacting with the adaptive agent. The agent attempts to mitigate the effects of both attentional tunnelling and information overload by reducing the visual clutter and perceptual

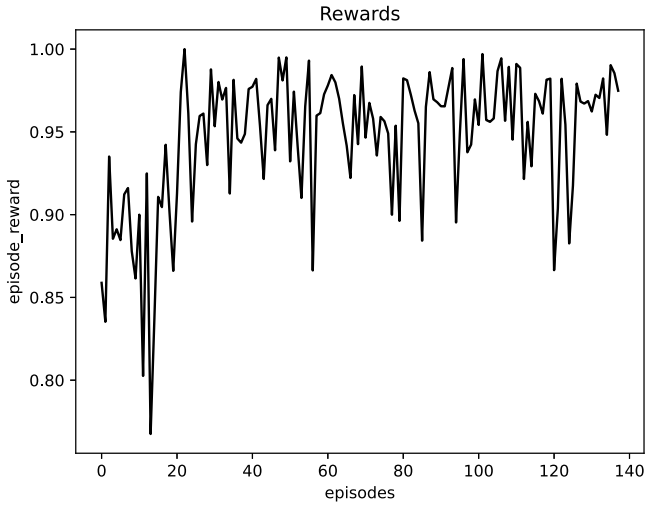

Fig. 3. Plot depicting the reward achieved by the agent over subsequent training episodes. As can be seen, the reward increases as the number of episode trained on increases; with occasional drops attributable to the probabilistic nature of the training environment.

load in the AR task. Because attentional tunnelling and information overload have been known to reduce task performance (speed and accuracy) (Kortschot and Jamieson, 2019; Gebhardt et al., 2019; Tatzgern et al., 2016), we expect users to have higher performance (speed and accuracy) when using the adaptive agent in comparison to users that do not use the adaptive agent. However, similar works using adaptive agents have failed to find a significant positive effect on task performance (Gebhardt et al., 2019; Lindlbauer et al., 2019; Koch et al., 2019). To better understand how the presence of an adaptive agent impacts user performance and behaviour while performing a sense-making task, we employed a mixed-methods approach (1) to assess users' performance (speed and accuracy) through quantitative means and (2) to understand users' interactions and experiences with the adaptive agent through qualitative feedback.

Although our system can support any task that involves clustering data points, we evaluated the system using a crime-solving sense-making task (see Section 2.2 for justification). The task we chose was Stanford and Stanford's murder mystery game (Stanford and Stanford, 1969). In this task, participants are presented with 31 text-based clues related to a fictional murder case and are tasked with identifying the name of the murderer, the weapon, the time, the place, and the motive behind the crime. Previous works have employed similar, sometimes more complex, crime-solving tasks to explore attention management in the context of reducing attention tunnelling in sense-making tasks (Goyal and Fussell, 2016, 2017). However, these tasks would require participants to continuously use the HoloLens head-mounted display for extended periods, which could lead to fatigue and postural discomfort (Knight and Baber, 2007). As such, the task we chose is shorter and can be completed in less than 60 min, but has been cited (Goyal and Fussell, 2017) and used (Zagermann et al., 2020) as a significantly complex and cognitively challenging sense-making task.

In line with prior work (Wozniak et al., 2016; Zagermann et al., 2020), we added 28 extra irrelevant clues to the murder mystery to increase its visual search complexity. Additionally, 1 clue was added to help users narrow down potential suspects, as some participants in our pilot test took longer than the battery capacity of the HoloLens 2 to verify their answers. This resulted in a total of 60 clues for our task. The number of additional clues we use is significantly larger than the number of clues added to the task in prior work. For example, Zagermann et al. (2020) added only 9 additional clues for the same murder mystery in their experiment to study a novel interaction

[3] Note that we do not carry over the training from current users to the following user to maintain consistency between users.

[4] Normalization of the rewards was based on the minimum and maximum rewards achieved by the agent and the random solver during the tests conducted, and the rewards achieved by the agent during its training phase.

Fig. 4. Images of (a) the experimental room with physical post-it notes arranged on a physical whiteboard and (b) a participant in an AR condition using the Hololens during the experiment.

method. However, our decision to add a larger number of clues was based on the need for our task to induce information overload, and high perceptual load. Specifically, we chose to include a total of 60 task items with similar visual features in our study as prior work has demonstrated both high perceptual load (Greene et al., 2017) and information overload (Gebhardt et al., 2019) using similar or less number of task items (see Table 1 for details of tasks used in previous works). See the Appendix for a list of all the 29 extra clues added to the original set.

All clues for the murder mystery consisted of a *body* and a *header*. The body contained the entirety of the clue, and the header was a single word representing the subject of the clue. The header for each clue was auto-generated through natural language processing, as described in Section 3.

We conducted the study in a laboratory setting with two wall-sized physical whiteboards (see Fig. 4(a), details presented in Appendix B). On one whiteboard, participants could find the questions they had to answer in the experimental task and the speech commands for interacting with the AR application. Participants were given markers that they could use to write their thoughts on the physical whiteboards. The experiment included three different conditions:

- **Paper:** The baseline condition where all clues for the murder mystery were presented on physical post-it notes. These notes were laminated and had a magnet attached to the back (Fig. 2(a)). Participants could move the notes around and attach them anywhere on two wall-sized physical whiteboards.
- **Unassisted AR:** Participants used our AR application to view virtual post-it notes containing the clues for the murder mystery task. Participants could move the notes around using hand gestures and place them anywhere in space, on the two physical wall-sized whiteboards or on the virtual boards that participants created using speech commands.
- **Assisted AR:** Similar to the Unassisted AR condition, participants used our AR application to view, manipulate and organize virtual post-it notes containing the clues for the murder mystery task. Our adaptive agent assisted participants in this condition by reducing or highlighting virtual notes based on the user's intentions and the notes the user considered relevant (see Section 3.2 for details).

5.1. Participants

Participants in the study expressed initial interest via an online form. The form asked participants about their experience and frequency of use of AR applications and technologies employing mid-air gestures. Participants were also asked about their experiences with games similar to murder mysteries and any work experience making sense of textual data (e.g. qualitative research, criminology, card sorting). Participants provided answers on a 5-point scale — Never, Less than monthly, Monthly, Weekly and Daily (Seabrook et al., 2020). We used stratified randomization based on these answers to balance participants amongst the three conditions of the study. Selected participants were then emailed to confirm their participation.

A total of 67 participants (32 men and 35 women) between the ages of 19 and 44 were recruited for the study. The data for 7 participants (5 men and 2 women) was discarded due to technical issues with the HoloLens, namely spatial mapping issues causing the rearrangement of virtual objects within the application or the HoloLens shutting down unexpectedly (the study could not be restarted as participants had already read through some clues of the murder mystery). This resulted in 20 participants for each condition.

5.2. Measures

We measured participants' task performance based on the time they took to provide all answers to the murder-mystery task (completion time) and the number of questions they answered correctly (accuracy). However, analysing completion time and accuracy separately would not be able to provide clear insights into how the different conditions affect overall performance. For example, a participant could complete the task quickly but find incorrect answers to the questions. Conversely, a participant could find all the correct answers but take too long to complete the task. Therefore, to enable the joint analysis of accuracy and speed, we use a combined measure called the Rate Correct Score (RCS) (Vandierendonck, 2017). RCS is defined as:

$$RCS = \frac{c}{\Sigma RT} \tag{4}$$

Where c is the number of correct responses and ΣRT is the sum of all response times for each question, i.e., the completion time for our task. RCS is interpreted as the number of correct responses per second of performing the task (Vandierendonck, 2017). As our task is fairly complex and requires a substantial amount of time, we present RCS in terms of *number of correct responses per minute of performing the task*.

We also measured participants' perceived task load via the NASA-TLX form (Hart and Staveland, 1988), which they completed after the sense-making task. We chose the NASA-TLX questionnaire as it considers both mental and physical demands. This is important, as in addition to mental demands, head-mounted AR devices have been known to cause neck fatigue (Bates and Istance, 2003) and postural discomfort (Cobb et al., 1999), which may affect physical task load. Additionally, NASA-TLX has been used extensively in prior work related to AR (Buchner et al., 2021), including visualizations (Biocca et al., 2006, 2007; Medenica et al., 2011) and adaptive agents (Lindlbauer et al., 2019). Lastly, the NASA-TLX questionnaire is applicable to all our conditions (Paper, Unassisted AR, Assisted AR), enabling us to perform comparisons—as opposed to using system-specific questionnaires applicable only to a subset of our conditions (e.g. Brooke et al., 1996; Körber, 2019).

We further employ semi-structured interviews to derive insights on condition-specific usability and user experiences. Specifically, the

interview focuses on the user's experience in solving a sense-making task using physical or AR tools, the benefits/challenges associated with the respective interactions and tool affordances, their prior experience with similar sense-making tasks, their strategies in solving the task, and their view on how the adaptive agent helped or hindered their performance (details in Appendix D).

5.3. Procedure

All procedures received approval from our institution's ethical review panel. Participants read a plain language statement of the study protocol and were asked to provide written consent to participate. Using a stratified-randomized design, we assigned participants to one of the three conditions (Paper, Unassisted AR or Assisted AR). Strata were used to control for participants' prior experience with mid-air gestural technology and problem-solving tasks (such as a qualitative researcher or criminologist). A participant's experience was determined through the online form they submitted when they volunteered for the study.

All participants were reminded that we would collect both their eye-tracking data and ego-centric video feed using either a Tobii Pro Glasses 2 Eye-Tracker (Paper condition) or the Microsoft HoloLens 2 (Unassisted AR and Assisted AR conditions). We then described the murder mystery task to the participants and directed them to the questions on the physical whiteboard. Participants were told to provide all 5 answers (not one at a time), in no specific order, to the researcher once they were satisfied with their conclusion and that they would receive an additional $5 (on top of $ 15 for participation) for getting all the answers correct. (This was used as an incentive at the start of the experiment, but all participants were rewarded with the total amount to ensure equity.) Participants were also informed that they would be timed from when they started the task to when they provided all 5 answers to the researcher and that they needed to complete the task as fast as possible.

Specific information was provided to participants based on the condition they were placed in:

- **Paper Condition**: Participants were first asked to wear a head-mounted eye-tracker (Tobii Pro Glasses 2). We then presented the physical post-it notes attached to a physical whiteboard (as shown in Fig. 4(a)). We instructed participants not to get close enough to read the clues before we started the experiment, as prior knowledge of the clues could potentially affect completion times. Participants were informed that they could move and organize the notes in any manner they chose. They could write on any physical whiteboard using the markers provided (4 different coloured markers and an eraser). The researcher started the timer once the participant had indicated that they were ready to start the task.
- **Unassisted AR Condition**: Participants were first asked to wear a Microsoft HoloLens 2 and calibrate the eye tracker included with the HoloLens. Participants then went through a 15-minute tutorial on the possible interactions in the AR system (See Section 3.1). They were told to practice the interactions until they felt confident in their use. We also informed participants that they could write on any physical whiteboard using the provided markers. When participants felt comfortable with the interactions and were clear on the task, they could start the experiment by facing a wall and using a speech command ('start'). This created 60 virtual post-it notes on the wall and started the timer for the task.
- **Assisted AR Condition**: Participants were provided with the same instructions and tutorial as with the AR condition. In addition, we described and showed a video demonstration of the different methods (Clustering, Folding and Outlining) that the adaptive agent used to help participants. To reduce uncertainty and misplaced expectations of the agent's capabilities — a core challenge identified in Human–AI interactions (Yang et al., 2020) — we provide detailed information regarding what the agent can and cannot do. Specifically, we explained that the adaptive agent did not know the answer to the sense-making task and only made decisions based on which notes the participant had been looking at, the notes the participant had deemed relevant by placing them on the relevant cluster boards and the similarity between the content of the different clues. We also explained that the agent could make better-informed decisions the more context it received, i.e., based on the number of post-it notes marked as relevant (for folding and outlining) and the relationship between notes on the different virtual boards (for clustering).

After finishing the task, participants completed the NASA-TLX questionnaire (Hart and Staveland, 1988) and took part in a 15 min long semi-structured interview. Participants were only given the correct answers to the sense-making task after completing the NASA-TLX form and interview.

The whole study (onboarding, experiment, and exit interview) took approximately 1 h, with 4/60 participants taking longer (ca. 15 min) due to longer completion times of the experiment task. Participants received a $20 gift voucher regardless of whether they got all answers correct.

5.4. Analysis

5.4.1. Quantitative data

We followed the statistical procedure suggested by Yatani (Yatani, 2014) to analyse our quantitative date.

As our collected RCS data deviated from a normal distribution based on the Shapiro–Wilk's test, we employed the Kruskal–Wallis test to determine if our conditions (Physical, AR and Assisted-AR) had a significant effect on the speed and accuracy of the sense-making task. Effect sizes for the Kruskal–Wallis tests were determined using the rstatix package in R (Kassambara, 2021). We further used a one-way ANOVA to investigate the cumulative task load of each condition based on our NASA-TLX data. Additionally, the effect of condition on each sub-scale of our NASA-TLX responses (mental demand, physical demand, temporal demand, performance, effort and frustration) was also investigated using an ANOVA or the Kruskal–Wallis test, depending on the normality of the data.

5.4.2. Qualitative data

We used the general inductive approach (Thomas, 2003) to analyse our interview responses. The primary researcher coded the interview data based on similar patterns observed in participant responses. Two researchers then independently analysed the data to better understand users' impressions on the tools used for the sense-making task (Paper versus AR) and assistance offered by the adaptive agent in the Assisted AR condition (see Appendix E). Video recordings for participants in the Assisted AR condition were also manually examined to verify self-reports related to user behaviour and interactions with the adaptive agent.

6. Results

6.1. Quantitative results

6.1.1. Rate correct score

Fig. 5 and Table 3 show the summary statistics for the rate correct score (RCS) data. The RCS is a combination of participants' accuracy (correct answers) and speed (completion time), corresponding to the number of correct answers per minute of task activity. The data shows that participants in all three conditions performed similarly in accuracy and speed. The average performance of participants in the Paper condition was slightly higher than both the Unassisted AR and Assisted

Table 3
Summary statistics of Rate Correct Score (RCS), presented as the number of correct answers per minute, grouped by Condition.

Condition	N	Mean (RCS)	SD (RCS)
Paper	20	0.162	0.066
Unassisted AR	20	0.138	0.060
Assisted AR	20	0.142	0.113

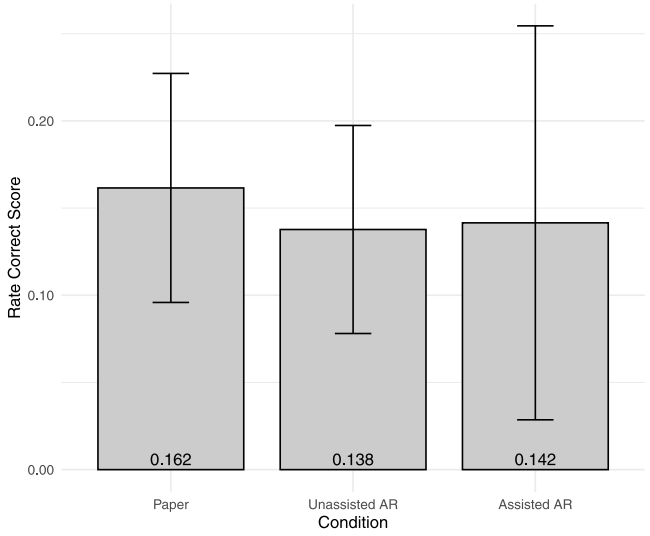

Fig. 5. Bar plot of RCS data grouped by condition. The data shows that average task performance was similar across all conditions. Standard deviation was similar between the Paper and Unassisted AR conditions, whereas the Assisted AR conditions showed much larger variation (approximately 2 times larger) in task performance.

AR conditions, while the average performance in the Assisted AR condition was slightly higher than the Unassisted AR condition. However, a Kruskal–Wallis test revealed no statistically significant differences between the conditions for the RCS score ($\chi^2 = 3.38,\ p = 0.18$). The effect size, $\eta^2 = 0.024$, indicates that Condition had a small effect on the number of correct answers per minute given by participants during the task (RCS). Although the standard deviation for task performance was similar between the Paper (0.066) and Unassisted AR (0.060) conditions, the Assisted-AR condition led to a much larger variation in task performance between participants (0.113).

6.1.2. Subjective workload

Table 4 shows the mean workload and standard deviation for each sub-scale of the NASA-TLX form as reported by participants, grouped by Condition. Data from the NASA-TLX for each participant was used to determine if there was a significant effect of condition on the overall workload. An ANOVA test revealed no significant effect of condition on overall workload ($F_{2,57} = 1.04,\ p = 0.36$), i.e., we did not find any evidence that the different conditions (Paper, Unassisted AR and Assisted AR) increased or decreased overall subjective workload significantly for participants. The size of the effect was also found to be small, $\eta^2 = 0.035$.

We further investigated responses for each sub-scale in the NASA-TLX forms. A Kruskal–Wallis test was used to determine the effect of Condition on physical demand. The test indicated a significant effect of condition on subjective physical demand ($\chi^2 = 9.64,\ p = 0.008$) with a medium effect size ($\eta^2 = 0.13$). Post-hoc analysis using a Mann–Whitney U test revealed that subjective physical demand in the Paper condition was significantly lower than in the Unassisted AR condition ($U = 108,\ Z = -2.503,\ p = 0.0113$). The size of the effect was medium, $r = 0.396$. A Mann–Whitney U test also revealed that subjective physical demand in the Paper condition was significantly lower than in the Assisted AR condition ($U = 98,\ Z = -2.769,\ p = 0.0049$). The size of the effect was medium, $r = 0.438$. This indicates that the task in the AR conditions was regarded to be more physically demanding than in the Paper condition.

A one-way ANOVA was used to determine the effect of the condition on subjective effort. The test indicated a significant effect of condition on subjective effort ($F_{2,57} = 3.177,\ p = .049$). The effect size was medium, $\eta^2 = 0.10$. Post-hoc analysis using a Welch Two-Sample T-test revealed that subjective effort was significantly higher in the Paper condition than in the Unassisted AR condition ($t(38) = 2.2659,\ p = 0.03$). The effect size was medium, Cohen's $d = 0.72$. This indicates that participants regarded the task to require more effort to complete in the Paper condition than in the Unassisted AR condition.

We found no significant effects of condition on all other sub-scales of the NASA-TLX responses (mental demand, temporal demand, performance and frustration).

6.2. Qualitative results

Our analysis of the interview responses and video recordings provides insights into three questions raised by the results of our quantitative analysis: (i) How did the tools and interactions offered in the AR conditions (Unassisted AR and Assisted AR) affect user behaviour and performance in comparison to the Paper condition? (ii) What factors influenced the larger variation in task performance in the Assisted AR condition? and (iii) What factors affected the use of the adaptive agent in the Assisted AR condition?

6.2.1. Effect of tools and interactions between the paper and AR conditions

To isolate the effects of our adaptive agent and the interactions available in the AR application, we first analysed participants' interview data for insights about how the different interactions offered in the Paper condition and AR conditions affected their behaviour and task performance.

Participants appreciated the interactions offered in the AR, reporting that the ability to interact with virtual objects outside their reach and organize the virtual objects in 3D space was useful. For example, P32 said *"It was easy because I could leave the notes in space but the physical ones could fall and I have to place them on actual walls"*. However, many participants in the Unassisted AR (9 participants) and Assisted AR (11 participants) conditions reported difficulties in interacting with the virtual objects. In contrast, no participant in the Paper condition reported any difficulties in interacting with the physical paper post-it notes.

Participants reported different reasons for why they struggled with the interactions offered in the AR application. Some participants reported that the gesture tracking was too sensitive, causing them to accidentally move objects. Others reported that depth perception made it difficult for them to interact with virtual objects. P22 said *"Sometimes I had to move back or forward to reach out (and grab notes)"*. There were also instances where the HoloLens 2 would not track participants' hands correctly if they looked away from their hands. P10 mentioned *"Dropped notes while turning and wasn't sure what happened"*..

Four participants in the Assisted AR condition said it was particularly frustrating to interact with notes in the Folded State due to their reduced size (see Fig. 2 b(ii)). For example, P46 said *"I think minimizing (folding) hindered my performance. (The notes became) too small to move"*.

Two participants in each of the Unassisted AR and Assisted AR conditions mentioned that the small field of view of the HoloLens 2 made it difficult for them to get an overview of all the notes. Five participants in the Unassisted AR and 7 participants in the Assisted AR mentioned that continued use of the HoloLens made it difficult for them to read the notes or caused fatigue.

Table 4
Mean and Standard deviation (within parenthesis) of the NASA-TLX scores for each sub-scale grouped by Condition. Mental demand for our task was seen to be higher than average in all three conditions. Effort was also seen to be higher than average in the Paper and Assisted AR conditions.

Condition	N	Mental demand	Physical demand	Temporal demand	Performance[6]	Effort	Frustration	Overall
Paper	20	55.0 (24.1)	18.0 (16.9)	38.8 (23.9)	48.5 (18.9)	62.8 (19.6)	28.2 (27.0)	41.9 (13.8)
Unassisted AR	20	57.8 (16.3)	33.0 (21.1)	43.0 (17.4)	39.5 (20.3)	48.5 (20.2)	22.5 (12.9)	40.7 (11.3)
Assisted AR	20	59.8 (18.5)	37.2 (24.3)	43.8 (19.3)	36.8 (20.8)	59.8 (16.6)	38.8 (23.3)	46.0 (11.5)

[6] Note that the *Performance* sub-scale is labelled from 'Perfect' to 'Failure' i.e., a lower score is associated with better performance, and vice-versa.

6.2.2. Variation in task performance in the assisted AR condition

The large variation in performance within the Assisted AR group prompted us to question why some participants performed better than others with the help of the adaptive agent. To better understand this variation, we performed a closer analysis of the responses and video recordings of participants with the lowest performance (up to the first quartile — bottom 5 participants with an average RCS = 0.040) and highest performance (fourth quartile and above — top 5 participants with an average RCS = 0.286) in the Assisted AR group. No obvious patterns in prior experience with mid-air gestures and problem solving tasks, as indicated by participants' responses to our screening survey, were detected between the lower performing and higher performing groups.

Based on the video and interview analysis of participants with performance measures in these quartiles, we found that four participants in the higher performance group (average RCS = 0.316) used the adaptive agent to filter, group and find notes. Folded notes, followed by notes clustered to different virtual boards, were reported and observed to be the most useful adaptations in filtering relevant from irrelevant information. In contrast, these participants reported that outlined notes were the least useful adaptation in helping to find relevant clues. P11 (RCS = 0.487) said *"(Folded notes were) good because you can still see the header and if you think the header is relevant, then you expand. It could have grabbed more of my attention because it was only the header and I think I would intuitively look at them first because it was cost-effective and I can just scan the header and decide if it's relevant or not"*.

Conversely, participants with lower performance ignored or actively undid the agent's actions. We observed that only 1 participant (RCS = 0.078) in the lower performance group occasionally attended to the clustered and outlined notes to filter clues. Another participant (RCS = 0, all answers were incorrect) focussed on the outlined notes at the start of the experiment but later abandoned this strategy. This participant was also observed attempting to filter irrelevant information by ignoring (or reading only the header of) notes in the *Folded State*. One other participant (RCS = 0, all answers were incorrect) said that the folded notes helped them to filter or quickly scan information, but video recordings of this participant showed otherwise. Four out of 5 participants (average RCS = 0.050) in the lower performance group actively undid folded notes, regardless of whether or not the header signified relevance. For example, P61 (RCS = 0.060) said *"I think I'd rather have not have it folding. I don't think it folded anything that was important but it would be better to see it (the note) fully rather than have to unfold it first"*.. Participants in the lower performance group were also observed to spend considerable time reading content of folded notes they expanded. This contrasts with the behaviour observed of participants in the higher performance group, who primarily expanded folded notes based on the header or to quickly confirm the note's relevance.

6.2.3. Factors affecting use of adaptive agent

We found that participants' individual perceptions of the adaptive agent were largely indicative of how they used the adaptations offered by the agent. While most participants (17 out of 20) who used the agent reported that at least one of the adaptations accurately helped in finding and filtering relevant notes, their perception of the adaptive agent's accuracy and capabilities — along with their own expectations and preferences regarding the agent's action — inhibited them from using the agent as intended. This in turn affected how well the agent could assist users in improving their task performance.

We identify three core factors affecting the use and perception of our adaptive agent based on participant responses and video recordings: (i) users' preferences of adaptations offered by the agent, (ii) users' trust in the agents actions and (iii) users' knowledge of the agents capabilities

User preferences. The first factor that appeared to affect participants' use of the agent was their preference towards particular adaptations and whether or not they chose to use these to complete the task. Participants expressed varying degrees of preference to the adaptations offered by the agent. *Clustering* was generally received favourably and *outlining* was largely ignored. Eight participants explicitly mentioned that clustering helped them find related content while only two mentioned that outlining relevant post-it notes helped them with the task. Remaining participants mentioned that they either ignored the *clustering* and/or *outlining*, or did not mention these adaptations during the interview.

In contrast, opinions surrounding the *folding* adaptation were varied. Four participants mentioned that the folding of notes helped them decide on whether a note was relevant or not. Another three said that folding of the notes both helped and harmed their performance as relevant information would occasionally be folded. Eight participants mentioned that they would unfold all folded notes as they preferred to see the complete content within notes.

However, some individuals expressed that they would prefer to avoid the interaction of unfolding notes altogether. P12 mentioned *"(I) did not quite like the collapsing of notes, (I) would prefer to see all notes and avoid the extra action to unfold a post-it note"*. The preference to not have any hidden content can be observed even in instances where the participant agrees that the agent is only hiding *irrelevant* information. P57 said *"The folding was actually correct, but I would still check just to make sure. And then I would go 'Okay, I actually didn't need that'"*..

One participant mentioned that they preferred to organize notes in 3D space rather than on the virtual boards. We also observed that five participants preferred to organized notes in 3D space in addition to using the virtual boards. For example, some placed notes closer to a relevant cluster board without attaching the note to the board if they were unsure the note's contents were related to that board. One participant did not use the virtual boards at all and exclusively used 3D space to organize their virtual notes.

Additionally, participants also mentioned that they would like explicit means of communicating their intentions with the agent in addition to the agent implicitly observing their gaze and actions to make decisions. For example, P66 said *"... the system was trying to find connections between my thoughts and the notes implicitly, but it would be useful to explicitly mention to the system that I was interested in 'Kelley' and 'weapon' and it would highlight all of those notes"*..

Trust in the adaptive agent. A second factor that influenced participants' behaviour was their level of trust towards the agent. Distrust towards the agent was primarily observed to manifest when the agent performed an action that the participant did not understand. These unexpected actions were observed to be more common at the beginning of our experiment, when the agent was still exploring what actions the user would reward/punish them for. Participants who observed the agent performing unexpected actions responded by either ignoring or undoing all the agents actions or by trying to understand how the agent worked and exploiting that knowledge.

As an example, one participant distrusted the agent and decided not to use the adaptations: P14 said *"(I would) prefer to read the text and did not trust the system. I would open all folded notes to read them"*.. In contrast, another participant observed that when the agent would fold relevant notes, irrelevant notes would be unfolded/outlined and vice-versa. This behaviour of the agent occurs in the earlier stages of the experiment when the user adopts the strategy of first discarding irrelevant notes without attaching many notes on the relevant cluster boards. Since, the agent observes the users gaze and the user is currently focused on irrelevant notes, the agent would outline irrelevant notes and fold relevant ones. P11 said *"Regardless of whether I trust the folding mechanism, If the folded note had a header that is not relevant, I would next look at a note that is not folded because it is presented differently. And vice-versa. If I see a folded title that is relevant, I would scan a couple other notes that are folded to see if I am missing anything that is folded"*.

Knowledge of the agent's capabilities. The third factor that affected participants' use of the agent was their knowledge of its capabilities. We observed that the lack of understanding of the adaptive agent's choice of action can lead users to ignore the agent's actions altogether. For instance, P1 says this about the outlining adaptation: *"the system did not help at all. Highlighting (Outlining) did not make sense and so I ignored"*.. Another example can be seen in P7's comments about the clustering adaptation: *"(I) used the colouring (clustering) first, (but) then found out it wasn't completely accurate and ignored"*..

Not understanding how the agent works can also influence the user's trust in the agent. P51 said that clustering helped overall but they had difficulty trusting the agent after the agent clustered notes into unexpected groups. *"Colouring (clustering) definitely helped. There was a lot of trust issues, especially in the beginning when I saw some colourings (clustered notes) made no sense. (I) didn't trust (the) agent when it folded notes and I would unfold it"*..

In contrast, P54 took note of our initial explanation of how the adaptive agent works and used the adaptations to find and categorize notes. *"I think the more I put things on the board, the more it helped. I thought it would have been better if I followed a strategy and placed specific topics on different boards ... At the end I had a couple of notes that was not placed on boards but the colouring helped me categorize them"*..

7. Discussion

This research aimed to better understand how adaptive agents can be used to mitigate attentional issues in AR sense-making tasks and the challenges associated with interacting with the adaptive agent. Our analysis of the speed and accuracy measures, along with the participant interviews, video recordings and subjective workload measures, led to three key findings.

First, according to our analysis of the NASA-TLX responses, participants rated the Unassisted AR condition to require less effort than the Paper condition. However, this benefit of using AR on subjective effort was not found between the Assisted AR and Paper condition, suggesting that the added complexity introduced by the adaptive agent offset the benefits of using AR on subjective effort. The analysis of the NASA-TLX responses also revealed that participants found the Paper condition to be less physically demanding than the AR conditions (Unassisted AR and Assisted AR). This was expected given that mid-air gestures are known to cause fatigue and discomfort (Hansberger et al., 2017). Despite the added physical demand and participant reports on how the novelty of the interactions and hardware limitations in the AR conditions negatively affected their task performance, we found no evidence to suggest that AR decreases task performance when compared to the Paper condition. This suggests that AR can be just as effective and more adaptable, through innovative means of interactions and data visualizations, in working with text-based sense-making data when compared to traditional paper-based methods.

Second, despite our positive technical evaluation and most participants (17 out of 20) in the Assisted AR condition reporting that at least one type of adaptation performed by the agent helped them filter information, we found that there was no significant improvement in the average task performance in the Assisted AR when compared to the Unassisted AR. However, we did observe a larger variation in task performance within the Assisted AR condition compared to the other conditions, suggesting that some participants were more successful in leveraging the capabilities of the adaptive agent to increase task performance. Our qualitative analysis indicates that participants with higher performance used the adaptations of the agent to find and filter relevant information effectively, while participants with lower performance ignored or actively undid the adaptations made by the agent. This suggests that participants who were more accepting of the agent's recommendations were more successful in drawing out helpful information from the adaptations.

Finally, we found that the user's behaviour towards the agent was primarily determined by their perception of the agent. We identified three key factors, based on interview responses and video recordings, that affected participants' use of the agent: (i) user preferences about the adaptations, (ii) trust in the agent's actions and (iii) knowledge of the agent's capabilities. The following subsection discusses the implications of these factors for the design of adaptive agents for AR applications.

7.1. User preferences and control over the agent

A key motivation for using reinforcement learning to create our adaptive agent was to enable the agent to cater towards individual users by observing individual users' decision strategies and user behaviours. This enabled our adaptive agent to perform different adaptations (reducing, clustering, and highlighting) to help participants find and filter relevant content in our AR sense-making task. While most participants found at least one adaptation helpful, the type of adaptation preferred by participants varied from individual to individual.

In our experiment, most participants who found the adaptations useful reported that *clustering* was helpful in finding related notes. This suggest that perceptual grouping using colours was effective in mitigating attention tunnelling (Wickens and Long, 1995), thus aiding visual search through increased response competition within groups (Fox, 1998; Baylis and Driver, 1992). In contrast, the *outlining* adaptation used in our system did not produce the intended effects of reducing perceptual load through visual guidance (Feit et al., 2020; Biggs et al., 2015). Most participants stated that they ignored the outlining effect. However, outlining as an adaptation to reduce perceptual load, may warrant future investigation. This is highlighted in our interviews which suggests that outlining has the potential to draw user attention but may be ineffective in tasks with high uncertainty — P11 said "The highlighting drew attention at first but I still had to read to see if the note was relevant or not"..

Lastly, opinions on the *folding* adaptation were largely mixed. Some participants reported that the folded notes helped in quickly filtering notes by enabling them scan the header alone, which made it more "cost-effective" than reading the full note. This suggests that adaptations aimed at reducing task-relevant information presented to users can mitigate information overload and lessen perceptual load (based on its operationalization (Macdonald and Lavie, 2011)). However, other participants stated that they would prefer not to have any hidden content in case they missed critical information. Based on our observations, this difference in opinion originates from the fact that, unlike the outlining and clustering, folded notes were difficult to ignore by participants that preferred not to have any content hidden. P11 reinforces our argument with the statement: *"Folding is the most prominent because it changes the shape of the notes"*.. In cases where an adaptation can be useful to some participants and harmful to others, it

would be beneficial for the user to request the agent only to use their preferred adaptations.

To handle adaptations that the user did not find favourable, we followed recent guidelines (Amershi et al., 2019) and implemented interactions enabling users to undo the agent's actions easily. This was implemented explicitly for the folding adaptation as it was the only adaptation that changed the information presented on a virtual-post-it note, i.e., by hiding the main body and displaying only the header of the note. Some participants, however, expressed that they would prefer to forgo the interaction and view the note's complete contents. This suggests the need to enable users to choose which adaptations the agent can perform or even to choose from alternative adaptations that achieve the same purpose (Feit et al., 2020). Creating adaptations to cater for different user preferences can be achieved by including potential users in the design of the agent's adaptations through participatory design methods (Schuler and Namioka, 1993; Bratteteig and Verne, 2018).

Participants were also observed to prefer different methods to organize virtual notes to solve the sense-making task. For example, some participants did not use the relevant cluster boards to organize their notes as intended but preferred to organize the notes in the 3D space. P26 said *"The most useful information about a note would be its position in space"*.. Not placing notes on the relevant cluster boards, however, limited the agent's information on what content the user deemed relevant and also removed feedback from the user on the different adaptations (see Section 3). This suggests the need to enable individual users to decide how they would prefer to indicate the relevance of notes to the agent.

Additionally, one participant commented that they would like to occasionally directly inform the agent what they were interested in and what the agent should do. In such instances, enabling users to communicate their intentions explicitly and needs to the agent would be helpful. For example, users could request the agent to outline all virtual notes related to the note they are gazing at.

7.2. Trust and knowledge of the AI agent's capabilities

Issues related to trust (Glass et al., 2008; Okamura and Yamada, 2020) and lack of knowledge of the AI's capabilities (Amershi et al., 2019) have been long-standing challenges for Human–AI interactions. Glass et al. argue that trusting and understanding the agent's actions is necessary for users to adopt and use these agents (Glass et al., 2008). Recent works around Human–AI interactions in HCI also stress the importance of building trust between the user and the AI agent (Amershi et al., 2019). These works provide guidelines to *avoid* trust issues with the AI agents (Amershi et al., 2019; Yang et al., 2020) by increasing user knowledge of the workings of the agent. They, however, do not discuss the *causes and effects* of trust (or distrust) on Human–AI interactions. This is significant as, despite following guidelines (Amershi et al., 2019) and explaining to our participants how, why and what our agent does prior to the experiment (see Section 5.3), some participants still distrusted the agent.

Distrusting the agent leads users to ignore the agent's actions and/or spend time undoing the agent's actions, thereby limiting (or outright eliminating) any potential benefits offered by the agent. In our study, participants' distrust of the agent primarily manifested through the expectation that the agent always outlined relevant notes and collapsed irrelevant notes *with respect to the sense-making task*. Our observations indicate that participants would distrust the agent if it did not perform actions in line with this expectation. As explained to the participants prior to the task, however, the agent does not know what is relevant to the sense-making task but performs actions based on *the participant's actions and eye tracking data*. This misalignment between the participants expectations and the agent's actual capabilities results in participant's not trusting future actions of the agent. This is consistent with previous works exploring attention cueing mechanisms, where users were reported to distrust the mechanism entirely after they encountered an incorrect or unexpected cue (Yeh et al., 2003).

Another contributing factor to users' trust of the agent was the user's awareness of how their actions and behaviour influenced the agent's actions. As an example, we observed that participants who adopted the strategy of finding and filtering out irrelevant notes at the start of the experiment would often observe the agent outlining irrelevant notes and collapsing relevant notes. Participants who did not understand that the agent was performing actions based on their intention of clearing out irrelevant notes would assume that the agent was inaccurately adapting the virtual content and would doubt the agent's capabilities. In contrast, participants who understood why the agent performed these actions were noted to take advantage of the adaptations (see Section 6.2.3).

The challenges associated with making the actions of AI understandable to users has led to multiple works focusing on explainable AI (XAI) (Das and Rad, 2020; Angelov et al., 2021; Liao et al., 2020). These works highlight the complexity and variability of user needs with understanding AI systems (Liao et al., 2020). In addition to challenges related to understanding AI systems, prior work on automated vehicles (Walker et al., 2023) highlights additional factors, such as experience with the system, the user's disposition, and situational context, that may influence trust. However, it is difficult to assume that factors influencing trust in automated vehicles would have similar effects on adaptive agents for sense-making tasks in AR, given the differences in consequences resulting from miscalibrated trust between the two scenarios. As such, further investigation is required to better understand overlapping factors affecting trust between adaptive systems used for different scenarios.

In addition to making the AI's actions more understandable (Liao et al., 2020) and implementing methods to undo the agent's actions (Amershi et al., 2019), our findings also highlight the need to enable users explicit means of providing feedback to the agent. In a study conducted by Wang et al. (2019), participants reported that they would be more inclined to trust the AI agent if they were presented with opportunities to provide their own feedback to the agent. However, an agent that constantly requires feedback could result in the user needing to expend more effort in providing input rather than focusing on the task.

Our agent was designed to receive feedback implicitly (see Section 3), thereby eliminating the need for the user to expend extra effort in providing explicit feedback to the agent. However, this eliminates the transparency of what feedback the agent is receiving, potentially reducing the user's trust in the agent (Wang et al., 2019) and leaving the user feeling ignored by the agent (Glass et al., 2008). Therefore, employing both implicit and explicit methods of providing feedback to the agent, could be beneficial (Li et al., 2021). This would enable the agent to collect implicit feedback, thereby reducing the burden on users needing to constantly provide input to the agent, while also enabling users to explicitly provide feedback when they feel the agent is not performing as intended.

7.3. Human-AI collaboration vs human-AI interaction

As suggested by Wang et al. interaction is not the same as collaboration (Wang et al., 2020). Collaboration requires mutual understanding of task goals and shared progress tracking of the task. Earlier concepts of collaborating with AI agents envisioned the human setting the goals and performing the evaluations while the agent assisted the human's efforts through routinizable work (Licklider, 1960). Recent works, however, have focused on designing autonomous agents with less opportunities for the user to provide feedback or input to the agent (Lindlbauer et al., 2019; Gebhardt et al., 2019; Feit et al., 2020).

Guidelines around Human–AI Interactions in HCI have attempted to address the consequences resulting from the lack of explicit means to provide input to adaptive agents (Amershi et al., 2019). These

guidelines include explaining the AI's capabilities to the user and implementing interactions to undo the AI's actions. However, our results demonstrate that these guidelines do not work for all users and can lead to users ignoring the agent's recommendations.

Enabling more direct means of communicating with the agent in addition to the agent's capabilities of gathering implicit data and feedback would help in establishing trust (Wang et al., 2019) with the agent and could lead to a more collaborative relationship between the user and the agent (Koch et al., 2019). The ability to explicitly inform the agent of the user's intention and provide explicit feedback to the agent would help the agent better understand the user's goals and, in turn, allow the agent to provide better assistance.

As an example case, the adaptive agent can collect implicit data and feedback from users and formulate actions accordingly without revealing these actions to the user. The user can then (i) toggle on or off all adaptations the agent offers based on current inferences, (ii) request the agent for specific adaptations, (iii) explicitly inform the agent of their intentions and/or (iv) force the agent to re-evaluate its actions by providing explicit feedback on its current actions.

Designing adaptive agents aimed at collaborating with the user requires a shift in the current focus from Human–AI Interactions to Human–AI Collaborations. While current guidelines focus on explaining the AI's actions and providing interactions to undo the actions that the agent has already performed, we highlight the need for more active measures in influencing the future actions of the agent.

8. Limitations and future work

Our adaptive agent was designed to mitigate attentional issues in AR sense-making tasks through the reduction of visual clutter and perceptual load. To evaluate the feasibility and performance of our agent, we conducted technical evaluations before running our main user study. Our technical evaluations demonstrate the agent's ability to learn a simulated user's behaviours and adapt to different sense-making tasks. However, we acknowledge that pre-evaluating the agent's performance with real users would further strengthen the arguments we present. Additionally, to determine the frequency with which our agent performs adaptations, we conducted a pilot study (see Appendix C) with three distinct time windows. While this enabled us to select a suitable time interval for adaptations to occur in our system, a more rigorous approach to optimize for the frequency of adaptations is beyond the scope of this paper and can be explored in future work.

Further, our work does not include the collection and analysis of raw gaze data. This was due to the lack of support in accessing raw gaze coordinates provided by the MRTKv2 API used for developing our AR application.[5] In order to collect raw user gaze data, gaze coordinates would have to be calculated and sent for data collection each frame. However, this caused significant performance issues with our application. In response, we opted to send aggregated gaze data, in ratios, needed by the agent every communication window (see Section 3). This necessity prevented us from collecting and analysing user gaze data.

The aim of the agent in mitigating attentional issues also limits the amount of information we could add to virtual content to explain the agent's actions within the AR application. For example, our agent could have displayed a numerical value on virtual post-it notes indicating relevance. However, this would add additional elements that the user would need to decipher, increasing perceptual load and defeating the purpose of our agent. A possible solution to this would be to fade or adjust transparency of virtual objects (Feit et al., 2020) based on the relevance of the content i.e., higher relevance equals less transparent and vice versa. Providing auditory explanations of the agent's actions could also help with issues related to user trust as previous works have demonstrated that visual perceptual load does not affect processing of information in the auditory stream (Tellinghuisen and Nowak, 2003). However, contradicting evidence indicating that high visual perceptual load could induce "inattentional deafness" (Macdonald and Lavie, 2011) — the phenomenon of missing *auditory* cues when attention is focused elsewhere. This suggest that users may not perceive the auditory explanations and as such further research is warranted before adopting auditory methods to explain the actions of an adaptive agent in AR.

Our work proposes the implementation of both implicit and explicit means to communicate with the adaptive agent for efficient collaboration. Future work can explore different ways to use both implicit and explicit user input. For example, implicit data such as walking or driving speed can inform adaptive agents for navigation applications in cars or smartphones about how many virtual elements to display without putting the user at risk of missing critical events. Explicit user commands in the same navigational application can then be used to request for specific information such distance to destination or navigational arrows. The agent in this case implicitly determines *how many* virtual elements is safe to display while the user determine *what* information they want displayed.

Future work can also explore how to balance between implicit and explicit interaction so that the agent can be used effectively. Adaptive agents need to be autonomous enough such that they do not turn into simple tools that users have to manually use or issue commands to. At the same time, the agent should allow for human input, so that users do not feel ignored and allow for better establishment of trust between the user and the agent.

9. Conclusion

This study shows how an adaptive agent aimed at reducing information overload and attentional tunnelling in AR sense-making tasks affects task performance and user behaviour. We found that users who were willing to accept the agent's recommendations and understand the different adaptations offered by the agent benefited more from the agent's assistance. We also determined key factors related to user preference, trust and knowledge of the adaptive agent that affects user behaviour towards our adaptive agent. We discuss the implications of these factors on task performance and highlight the need to design for better collaboration between users and adaptive agents.

CRediT authorship contribution statement

Brandon Victor Syiem: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review & editing. **Ryan M. Kelly:** Formal analysis, Writing – review & editing. **Tilman Dingler:** Writing – review & editing, Supervision. **Jorge Goncalves:** Conceptualization, Supervision, Writing – review & editing. **Eduardo Velloso:** Conceptualization, Funding acquisition, Methodology, Supervision, Writing – original draft, Writing – review & editing.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Eduardo Velloso reports financial support was provided by Australian Research Council.

Data availability

The authors do not have permission to share data.

[5] https://learn.microsoft.com/en-us/windows/mixed-reality/mrtk-unity/mrtk2/features/input/eye-tracking/eye-tracking-eye-gaze-provider?view=mrtkunity-2022-05

Acknowledgements

We thank our participants for their time. This research was supported by the Melbourne Research Scholarship, Australia provided by the University of Melbourne. Eduardo Velloso is the recipient of an Australian Research Council Discovery Early Career Award (Project Number: DE180100315) funded by the Australian Government.

Appendix A. Additional clues in the sense-making task

A.1. Relevant clues to help users

1. Miss Smith said Mr. Jones did not see very well.

A.2. Irrelevant clues to hinder users

1. Mr. Kelley went for a walk at 2:00pm on the day before his body was found.
2. A large number of robberies were reported in the area.
3. The elevator man said that he often witnessed Miss Smith and Mr. Kelley arguing.
4. Miss Smith enjoyed taking long walks in the evenings.
5. Mr. Jones was a gun enthusiast.
6. Mr. Kelley visited the Taj Mahal in India in the 1980's.
7. The elevator man was named Rob James.
8. Construction of the Taj Mahal was completed in 1643.
9. Miss Smith had a cat named "Mittens" who liked to sleep under the coffee table.
10. The largest fish in the world is the Whale Shark.
11. World War 2 lasted from 1939 to 1945.
12. No number from 1 to 999 includes the letter "a" in its word form.
13. The opposite sides of a die will always add up to seven.
14. A Greek-Canadian man invented the "Hawaiian" pizza.
15. Cats can't taste sweet things because of a genetic defect.
16. A group of hippos is called a "bloat".
17. A "jiffy" is about one trillionth of a second.
18. Dragonflies have six legs but can't walk.
19. Apple seeds contain cyanide.
20. A frigate bird can sleep while it flies.
21. Jupiter is twice as massive as all the other planets combined.
22. Your body contains about 100,000 miles of blood vessels.
23. The inventor of Pringles is buried in a Pringles can.
24. The largest scrambled eggs ever made weighed nearly 3.5 tons.
25. Octopuses and squid have three hearts.
26. The first email was sent by Ray Tomlinson to himself in 1971.
27. Dolphins give each other names.
28. Marie Curie's 100-year-old belongings are still radioactive.

Appendix B. Laboratory and apparatus

Here we detail the dimension of the experiment room, whiteboards (physical and virtual), and post-it notes (physical and virtual) that we used for the study. All measurements are in millimeters (mm).

1. **Experiment room:** The room measured 5450 mm in length, 3570 mm in width, and 2580 mm in height. Room walls either consisted of whiteboards or fabric panels of equal dimensions, with one side consisting of a door and a one way mirror. The door and mirror together measured 3570 mm and the mirror covered the space from the ceiling halfway to the floor (1300 mm), with fabric continuing from the end of the mirror to the floor. One wall along the length of the room consisted of 5 fabric panels, with the opposite side containing 3 whiteboard panels (making up 1 large continuous whiteboard) between 2 fabric panels. The last wall, along the width and opposite the door, consisted of 3 whiteboard panels (making up another large continuous whiteboard).
2. **Physical whiteboard:** Each whiteboard panel measured 1090 mm in width and 2470 mm in height.
3. **Physical post-it note:** Each physical post-it note used in the Physical condition measured 93 mm in length and 93 mm in width.
4. **Virtual whiteboard:** Virtual whiteboards measured 651 mm in width and 511 mm in length.[6]
5. **Virtual post-it note:** Each virtual post-it note used in the Unassisted AR and Assisted AR condition measured 93 mm in length and 93 mm in width. This is consistent with the dimensions of the physical post-it notes.

Appendix C. Pilot study

Prior to running our study described in Section 5, we ran a pilot test with 6 participants to test and refine our system, and the task used for our experiment. In the pilot test, we tested for the feasibility of completing the chosen task in the Assisted AR condition. We used the same murder mystery task used in our final experiment but included only the 28 irrelevant clues detailed in Appendix A.2. The number of additional clues was chosen to ensure that our task had a similar, or more, number of task relevant items with similar features to induce high perceptual load. Specifically, we chose to have more than 51 total task related items, based on the study by Greene et al. (2017) that *explicitly* focused on the effects of perceptual load on task performance, between conditions with 51 and 13 additional task items (see Table 1 for details).

Additionally, the agent used for the pilot test initially performed adaptations every 3 s (communication window). This time window was chosen based on average gaze fixation duration on words during reading reported in prior work (Liversedge et al., 1998; Yang, 2009; Staub et al., 2010) (approximately 200 ms to 300 ms) along with the average number of words in each sentence of our task (677 words between the 59 clues, giving us an average of 11.47 word per clue) i.e., $250 \times 11.47 = 2867.5$ ms. We followed the same procedure as detailed in Section 5. However, our final interview for the pilot was focused on user feedback on the adaptions and task difficulty. Participant data for the pilot study was not recorded, and interview responses and observational notes were used only as immediate feedback for necessary changes.

Our first 2 participants commented that they ignored the adaptations performed by the agent due to the frequency with which these actions were performed. They also reported that the task felt too difficult and frustrating. However, it was not clear if the difficulty was due to the task itself or caused by the excessive adaptions performed by the agent. This prompted us to reflect on our decision of a 3 s communication widow that was based solely on reading times and did not account for the necessary search for post-it notes to read. We therefore decided to conduct further testing with a 5 and 7 s window, with 2 participants in each condition.

We found that participants in the 5 s window did not comment on the frequency of the agent's actions. However, participants in the 7 s window condition questioned whether the agent was lagging behind. In addition, 2 participants (1 in the 3 s and the other in the 5 s condition) said that they had to guess a few answers for the task as the AR device ran out of battery. This prompted us to add 1 additional clue that assisted users in solving the task for the final study (Appendix A.2).

[6] Note: virtual post-it notes could be attached to either virtual or physical whiteboards. Virtual whiteboards were necessary for the agent to determine relevance of attached post-it notes

Appendix D. Semi-structured interview questions

The post-study semi-structured interview that we conducted focused on: the user's experience with solving the sense-making task with the given tools (Paper, Unassisted AR, Assisted AR), the challenges and benefits in using the tools, the user's prior experiences with similar tasks and how prior experiences may have shaped their progression of the task, and the strategies — and changes in strategy — that users employed during the task. These questions aimed to gain a broader understanding of the user experience, their chosen strategies, and the influence of the given tools in solving the sense-making task. We asked users additional questions related to the impact of the adaptive agent in the Assisted AR condition. We present the high-level questions that we asked during the interview below:

D.1. Questions asked in all conditions

1. How was your experience in solving the murder mystery/sense-making task using the physical/virtual post-it notes and white-boards?
2. Did you find anything challenging[7] or useful in using the given tools for solving the task? (elaborate on why? why not?)
3. Have you ever solved similar sense-making problems before? (elaborate)
4. Is there a strategy that you used to approach the problem? (follow up examples - change of strategy? focus on mystery as a whole or individual questions? why chosen strategy?)

D.2. Questions added for the assisted AR condition

1. Did you use any of the suggestions made by the adaptive agent? (elaborate on why? why not?)
2. Could you elaborate on what you think about each of the adaptations (clustering using colours, outlining or folding) in relation to their role in supporting/hindering your task? (if support - how did it support the task and to what extent? if hinder - what problems/challenges did you notice with the adaptations?)

Appendix E. Categories developed during our analysis of the interviews

We developed 85 codes from our interview data, and further group these codes into 7 categories: Answering Strategies, Conceptual Organization, Spatial Organization, Elements Affecting Task Performance, Tools and Interactions, Effects of Adaptations, Other (engagement, difficulties, and comparisons between virtual and physical).

Appendix F. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.ijhcs.2024.103324.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X., 2015. TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org. URL https://www.tensorflow.org/.

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P.N., Inkpen, K., Teevan, J., Kikin-Gil, R., Horvitz, E., 2019. Guidelines for human-AI interaction. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Andrews, C., Endert, A., North, C., 2010. Space to think: Large high-resolution displays for sensemaking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '10, Association for Computing Machinery, New York, NY, USA, pp. 55–64. http://dx.doi.org/10.1145/1753326.1753336.

Angelov, P.P., Soares, E.A., Jiang, R., Arnold, N.I., Atkinson, P.M., 2021. Explainable artificial intelligence: an analytical review. Wiley Interdiscip. Rev.: Data Min. Knowl. Discov. 11 (5), e1424.

Ayers, J.W., Leas, E.C., Dredze, M., Allem, J.-P., Grabowski, J.G., Hill, L., 2016. Pokémon GO—A new distraction for drivers and pedestrians. JAMA Intern. Med. 176 (12), 1865–1866. http://dx.doi.org/10.1001/jamainternmed.2016.6274, arXiv:https://jamanetwork.com/journals/jamainternalmedicine/articlepdf/2553331/ild160053.pdf.

Bates, R., Istance, H.O., 2003. Why are eye mice unpopular? A detailed comparison of head and eye controlled assistive technology pointing devices. Univers. Access Inf. Soc. 2, 280–290.

Baylis, G.C., Driver, J., 1992. Visual parsing and response competition: The effect of grouping factors. Percept. Psychophys. 51 (2), 145–162.

Biggs, A.T., Kreager, R.D., Davoli, C.C., 2015. Finding a link between guided search and perceptual load theory. J. Cogn. Psychol. 27 (2), 164–179.

Biocca, F., Owen, C., Tang, A., Bohil, C., 2007. Attention issues in spatial information systems: Directing mobile users' visual attention using augmented reality. J. Manage. Inf. Syst. 23 (4), 163–184.

Biocca, F., Tang, A., Owen, C., Xiao, F., 2006. Attention funnel: Omnidirectional 3D cursor for mobile augmented reality platforms. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. CHI '06, ACM, New York, NY, USA, pp. 1115–1122. http://dx.doi.org/10.1145/1124772.1124939, URL http://doi.acm.org/10.1145/1124772.1124939.

Bratteteig, T., Verne, G., 2018. Does AI make PD obsolete? Exploring challenges from artificial intelligence to participatory design. In: Proceedings of the 15th Participatory Design Conference: Short Papers, Situated Actions, Workshops and Tutorial - Volume 2. PDC '18, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3210604.3210646.

Brooke, J., et al., 1996. SUS-a quick and dirty usability scale. Usability Eval. Ind. 189 (194), 4–7.

Buchner, J., Buntins, K., Kerres, M., 2021. A systematic map of research characteristics in studies on augmented reality and cognitive load. Comput. Educ. Open 2, 100036.

Cartwright-Finch, U., Lavie, N., 2007. The role of perceptual load in inattentional blindness. Cognition 102 (3), 321–340. http://dx.doi.org/10.1016/j.cognition.2006.01.002, URL https://www.sciencedirect.com/science/article/pii/S0010027706000205.

Casas, N., 2017. Deep deterministic policy gradient for urban traffic light control. arXiv preprint arXiv:1703.09035.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., John, R.S., Constant, N., Guajardo-Céspedes, M., Yuan, S., Tar, C., et al., 2018. Universal sentence encoder. arXiv preprint arXiv:1803.11175.

Chollet, F., et al., 2015. Keras. https://keras.io.

Cobb, S.V., Nichols, S., Ramsey, A., Wilson, J.R., 1999. Virtual reality-induced symptoms and effects (VRISE). Presence: Teleoperators Virt. Environ. 8 (2), 169–186.

Das, A., Rad, P., 2020. Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv preprint arXiv:2006.11371.

De Graaf, M., Allouch, S.B., Van Diik, J., 2017. Why do they refuse to use my robot?: Reasons for non-use derived from a long-term home study. In: 2017 12th ACM/IEEE International Conference on Human-Robot Interaction. HRI, IEEE, pp. 224–233.

Feit, A.M., Vordemann, L., Park, S., Berube, C., Hilliges, O., 2020. Detecting relevance during decision-making from eye movements for UI adaptation. In: ACM Symposium on Eye Tracking Research and Applications. In: ETRA '20 Full Papers, Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3379155.3391321.

Fox, E., 1998. Perceptual grouping and visual selective attention. Percept. Psychophys. 60 (6), 1004–1021.

Gebhardt, C., Hecox, B., van Opheusden, B., Wigdor, D., Hillis, J., Hilliges, O., Benko, H., 2019. Learning cooperative personalized policies from gaze data. In: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. UIST '19, Association for Computing Machinery, New York, NY, USA, pp. 197–208. http://dx.doi.org/10.1145/3332165.3347933.

Glass, A., McGuinness, D.L., Wolverton, M., 2008. Toward establishing trust in adaptive agents. In: Proceedings of the 13th International Conference on Intelligent User Interfaces. IUI '08, Association for Computing Machinery, New York, NY, USA, pp. 227–236. http://dx.doi.org/10.1145/1378773.1378804.

Goyal, N., Fussell, S.R., 2016. Effects of sensemaking translucence on distributed collaborative analysis. In: Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing. CSCW '16, Association for Computing Machinery, New York, NY, USA, pp. 288–302. http://dx.doi.org/10.1145/2818048.2820071.

[7] We followed up with questions related to motion sickness, depth perception, gesture recognition, etc. for the AR conditions

Goyal, N., Fussell, S.R., 2017. Intelligent interruption management using electro dermal activity based physiological sensor for collaborative sensemaking. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 1 (3), http://dx.doi.org/10.1145/3130917.

Greene, C.M., Murphy, G., Januszewski, J., 2017. Under high perceptual load, observers look but do not see. Appl. Cogn. Psychol. 31 (4), 431–437.

Hansberger, J.T., Peng, C., Mathis, S.L., Areyur Shanthakumar, V., Meacham, S.C., Cao, L., Blakely, V.R., 2017. Dispelling the gorilla arm syndrome: The viability of prolonged gesture interactions. In: Lackey, S., Chen, J. (Eds.), Virtual, Augmented and Mixed Reality. Springer International Publishing, Cham, pp. 505–520.

Hart, S.G., Staveland, L.E., 1988. Development of NASA-TLX (task load index): Results of empirical and theoretical research. In: Hancock, P.A., Meshkati, N. (Eds.), Human Mental Workload. In: Advances in Psychology, vol. 52, North-Holland, pp. 139–183. http://dx.doi.org/10.1016/S0166-4115(08)62386-9, URL https://www.sciencedirect.com/science/article/pii/S0166411508623869.

Kassambara, A., 2021. rstatix. https://rpkgs.datanovia.com/rstatix/index.html.

Knight, J.F., Baber, C., 2007. Effect of head-mounted displays on posture. Hum. Factors 49 (5), 797–807.

Koch, J., Lucero, A., Hegemann, L., Oulasvirta, A., 2019. May AI? Design ideation with cooperative contextual bandits. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–12. http://dx.doi.org/10.1145/3290605.3300863.

Körber, M., 2019. Theoretical considerations and development of a questionnaire to measure trust in automation. In: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20. Springer, pp. 13–30.

Kortschot, S.W., Jamieson, G.A., 2019. Classification of attentional tunneling through behavioral indices. Hum. Factors 0018720819857266.

Lavie, N., 1995. Perceptual load as a necessary condition for selective attention. J. Exp. Psychol. [Hum. Percept.] 21 (3), 451.

Li, Z., Shi, L., Cristea, A.I., Zhou, Y., 2021. A survey of collaborative reinforcement learning: Interactive methods and design patterns. In: Designing Interactive Systems Conference 2021. Association for Computing Machinery, New York, NY, USA, pp. 1579–1590. http://dx.doi.org/10.1145/3461778.3462135.

Liao, Q.V., Gruen, D., Miller, S., 2020. Questioning the AI: informing design practices for explainable AI user experiences. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–15.

Licklider, J.C.R., 1960. Man-computer symbiosis. IRE Trans. Hum. Factors Electron. HFE-1 (1), 4–11. http://dx.doi.org/10.1109/THFE2.1960.4503259.

Lindlbauer, D., Feit, A.M., Hilliges, O., 2019. Context-aware online adaptation of mixed reality interfaces. In: Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology. UIST '19, Association for Computing Machinery, New York, NY, USA, pp. 147–160. http://dx.doi.org/10.1145/3332165.3347945.

Liversedge, S.P., Paterson, K.B., Pickering, M.J., 1998. Eye movements and measures of reading time. In: Eye Guidance in Reading and Scene Perception. Elsevier, pp. 55–75.

Macdonald, J.S., Lavie, N., 2011. Visual perceptual load induces inattentional deafness. Atten. Percept. Psychophys. 73 (6), 1780–1789. http://dx.doi.org/10.3758/s13414-011-0144-4.

Medenica, Z., Kun, A.L., Paek, T., Palinko, O., 2011. Augmented reality vs. street views: a driving simulator study comparing two emerging navigation aids. In: Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services. pp. 265–274.

Okamura, K., Yamada, S., 2020. Adaptive trust calibration for human-AI collaboration. PLoS One 15 (2), e0229132. http://dx.doi.org/10.1371/journal.pone.0229132.

Pirolli, P., Card, S., 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In: Proceedings of International Conference on Intelligence Analysis, Vol. 5. McLean, VA, USA, pp. 2–4.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, URL https://arxiv.org/abs/1908.10084.

Schick, A.G., Gordon, L.A., Haka, S., 1990. Information overload: A temporal approach. Account. Organ. Soc. 15 (3), 199–220. http://dx.doi.org/10.1016/0361-3682(90)90005-F, URL https://www.sciencedirect.com/science/article/pii/036136829090005F.

Schuler, D., Namioka, A., 1993. Participatory Design: Principles and Practices. CRC Press.

Seabrook, E., Kelly, R., Foley, F., Theiler, S., Thomas, N., Wadley, G., Nedeljkovic, M., 2020. Understanding how virtual reality can support mindfulness practice: mixed methods study. J. Med. Internet Res. 22 (3), e16106. http://dx.doi.org/10.2196/16106.

Stanford, G., Stanford, B.D., 1969. Learning Discussion Skills Through Games. ERIC.

Staub, A., White, S.J., Drieghe, D., Hollway, E.C., Rayner, K., 2010. Distributional effects of word frequency on eye fixation durations. J. Exp. Psychol. [Hum. Percept.] 36 (5), 1280.

Subramonyam, H., Drucker, S.M., Adar, E., 2019. Affinity lens: Data-assisted affinity diagramming with augmented reality. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3290605.3300628.

Syiem, B.V., Kelly, R.M., Goncalves, J., Velloso, E., Dingler, T., 2021. Impact of task on attentional tunneling in handheld augmented reality. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, http://dx.doi.org/10.1145/3411764.3445580.

Syiem, B.V., Kelly, R.M., Velloso, E., Goncalves, J., Dingler, T., 2020. Enhancing visitor experience or hindering docent roles: Attentional issues in augmented reality supported installations. In: 2020 IEEE International Symposium on Mixed and Augmented Reality. ISMAR, pp. 279–288. http://dx.doi.org/10.1109/ISMAR50242.2020.00053.

Tatzgern, M., Orso, V., Kalkofen, D., Jacucci, G., Gamberini, L., Schmalstieg, D., 2016. Adaptive information density for augmented reality displays. In: 2016 IEEE Virtual Reality. VR, pp. 83–92. http://dx.doi.org/10.1109/VR.2016.7504691.

Tellinghuisen, D.J., Nowak, E.J., 2003. The inability to ignore auditory distractors as a function of visual task perceptual load. Percept. Psychophys. 65 (5), 817–828.

Thomas, D.R., 2003. A General Inductive Approach for Qualitative Data Analysis. Citeseer.

Treisman, A., 1982. Perceptual grouping and attention in visual search for features and for objects.. J. Exp. Psychol. [Hum. Percept.] 8 (2), 194.

Vandierendonck, A., 2017. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. Behav. Res. Methods 49 (2), 653–673. http://dx.doi.org/10.3758/s13428-016-0721-5.

Wagner-Greene, V.R., Wotring, A.J., Castor, T., Kruger, J., Mortemore, S., Dake, J.A., 2017. Pokémon GO: Healthy or harmful? Am. J. Public Health 107 (1), 35. http://dx.doi.org/10.2105/AJPH.2016.303548.

Walker, F., Forster, Y., Hergeth, S., Kraus, J., Payre, W., Wintersberger, P., Martens, M., 2023. Trust in automated vehicles: constructs, psychological processes, and assessment. Front. Psychol. 14, 1279271.

Wang, D., Churchill, E., Maes, P., Fan, X., Shneiderman, B., Shi, Y., Wang, Q., 2020. From human-human collaboration to human-AI collaboration: Designing ai systems that can work together with people. In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. In: CHI EA '20, Association for Computing Machinery, New York, NY, USA, pp. 1–6. http://dx.doi.org/10.1145/3334480.3381069.

Wang, D., Weisz, J.D., Muller, M., Ram, P., Geyer, W., Dugan, C., Tausczik, Y., Samulowitz, H., Gray, A., 2019. Human-AI collaboration in data science: Exploring data scientists' perceptions of automated AI. Proc. ACM Hum.-Comput. Interact. 3 (CSCW), http://dx.doi.org/10.1145/3359313.

Wickens, C.D., Alexander, A.L., 2009. Attentional tunneling and task management in synthetic vision displays. Int. J. Aviat. Psychol. 19 (2), 182–199. http://dx.doi.org/10.1080/10508410902766549.

Wickens, C.D., Long, J., 1995. Object versus space-based models of visual attention: Implications for the design of head-up displays.. J. Exp. Psychol. Appl. 1 (3), 179.

Wickens, C.D., Martin-Emerson, R., Larish, I., 1993. Attentional tunneling and the head-up display.

Wozniak, P., Goyal, N., Kucharski, P.a., Lischke, L., Mayer, S., Fjeld, M., 2016. RAMPARTS: Supporting sensemaking with spatially-aware mobile interactions. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 2447–2460. http://dx.doi.org/10.1145/2858036.2858491.

Yang, S.-N., 2009. Effects of gaze-contingent text changes on fixation duration in reading. Vis. Res. 49 (23), 2843–2855.

Yang, Q., Steinfeld, A., Rosé, C., Zimmerman, J., 2020. Re-examining whether, why, and how human-AI interaction is uniquely difficult to design. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13.

Yatani, K., 2014. Statistical methods for HCI research. https://yatani.jp/teaching/doku.php?id=hcistats:start.

Yeh, M., Merlo, J.L., Wickens, C.D., Brandenburg, D.L., 2003. Head up versus head down: The costs of imprecision, unreliability, and visual clutter on cue effectiveness for display signaling. Hum. Factors 45 (3), 390–407. http://dx.doi.org/10.1518/hfes.45.3.390.27249, PMID: 14702991.

Zagermann, J., Pfeil, U., von Bauer, P., Fink, D., Reiterer, H., 2020. "It's in my other hand!" – Studying the interplay of interaction techniques and multi-tablet activities. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, NY, USA, pp. 1–13. http://dx.doi.org/10.1145/3313831.3376540.