Can AR Embedded Visualizations Foster Appropriate Reliance on AI in Spatial Decision-Making? A Comparative Study of AR X-Ray vs. 2D Minimap

Xianhao Carton Liu
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
liu03008@umn.edu

Difan Jia
Department of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
bobby003@umn.edu

Tongyu Nie
Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota, USA
nie00035@umn.edu

Evan Suma Rosenberg
Department of Computer Science & Engineering
University of Minnesota
Minneapolis, Minnesota, USA
suma@umn.edu

Victoria Interrante
Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota, USA
interran@cs.umn.edu

Chen Zhu-Tian
Department of Computer Science and Engineering
University of Minnesota-Twin Cities
Minneapolis, Minnesota, USA
ztchen@umn.edu

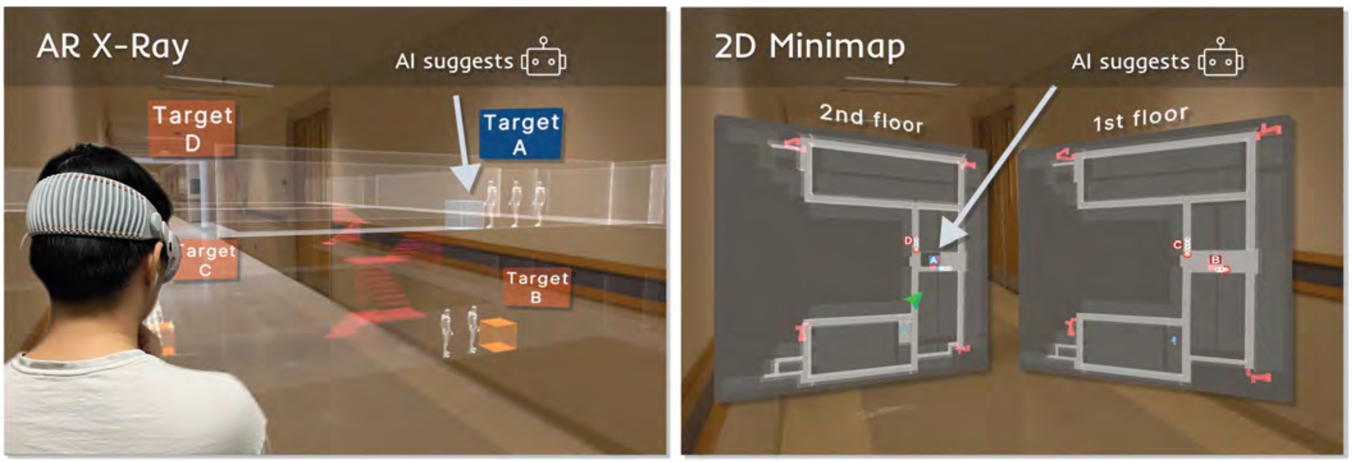

Figure 1: A participant wearing an Apple Vision Pro performs spatial decision-making tasks using two different visualizations: AR X-ray (Left) and 2D Minimap (Right). Both visualizations present AI-suggested targets, but the AR X-ray approach embeds information directly into the participant's field of view, while the 2D Minimap offers a top-down abstraction.

Abstract

Artificial Intelligence (AI) and indoor sensing increasingly support decision-making in spatial environments. However, traditional visualization methods impose a substantial mental workload when viewers translate this digital information into real-world spaces, leading to inappropriate reliance on AI. Embedded visualizations in Augmented Reality (AR), by integrating information into physical environments, may reduce this workload and foster more appropriate reliance on AI. To assess this, we conducted an empirical study (N = 32) comparing an AR embedded visualization (X-ray) and 2D Minimap in AI-assisted, time-critical spatial target selection tasks. Surprisingly, evidence shows that the embedded visualization led to greater inappropriate reliance on AI, primarily as over-reliance, due to factors like perceptual challenges, visual proximity illusions, and highly realistic visual representations. Nonetheless, the embedded visualization showed benefits in spatial mapping. We conclude by discussing empirical insights, design implications, and directions for future research on human-AI collaborative decision in AR.

CCS Concepts

• **Human-centered computing** → **Empirical studies in visualization**; **Empirical studies in HCI**; • **Computing methodologies** → **Machine learning**.

Keywords

Augmented Reality, Decision-Making, Human-AI Collaboration, Embedded Visualizations

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3790710

ACM Reference Format:
Xianhao Carton Liu, Difan Jia, Tongyu Nie, Evan Suma Rosenberg, Victoria Interrante, and Chen Zhu-Tian. 2026. Can AR Embedded Visualizations Foster Appropriate Reliance on AI in Spatial Decision-Making? A Comparative Study of AR X-Ray vs. 2D Minimap. In *Proceedings of the*

2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3772318.3790710

1 Introduction

From high-stakes emergency evacuations [60, 99] and crisis management [48], to everyday situations like hurrying to a classroom or navigating an unfamiliar building under time pressure, people often have to rapidly choose among multiple spatial targets such as exits, areas to secure, and hallways. We refer to these decision-making tasks, which require integrating information about spatial environments with external data, as *Spatial Decision-Making*. Making a poor choice can waste time and resources, and in the worst cases, endanger lives [33]. Recent advancements in indoor sensing and artificial intelligence (AI) offer new avenues for supporting such decisions by visualizing real-time situational data and AI suggestions in a spatial context [53, 69].

However, the way this information is presented can significantly impact user behavior and decision quality [26]. Existing methods [54, 102] commonly employ 2D maps to display data and AI suggestions, requiring users to mentally map and connect these digital cues onto physical spaces. According to spatial cognition research [44], this process, known as *reference frame translation*, can impose a substantial cognitive load. Under time pressure, users may struggle to critically assess AI outputs, resulting in *inappropriate reliance* [85]: either trusting flawed AI suggestions or rejecting correct ones. This concern is especially pressing in emergency contexts, where every second counts.

Augmented Reality (AR) promises a more intuitive approach by directly visualizing data and AI suggestions onto the physical environment. These visualizations, often referred to as *embedded visualizations* [97, 109], can reduce cognitive load by eliminating the need for constant reference frame translation. Intuitively, embedded visualizations might foster more deliberate and appropriate reliance on AI. Yet, whether embedded visualizations genuinely mitigates inappropriate reliance on AI in spatial decision-making remains unclear, and little existing work has examined how embedded visualizations in AR affects human-AI reliance in spatial tasks. As AR and AI continue to mature and increasingly integrate [34], this work aims to provide researchers and practitioners with initial evidence on how embedded visualizations impact human-AI reliance in spatial decision-making.

To this end, we present an empirical study (N=32) comparing an AR X-ray (Figure 1 Left) against a 2D Minimap (Figure 1 Right) in a time-pressured, AI-assisted spatial decision task. Following previous work on AR visual systems for indoor navigation [69], we selected the X-ray view as our AR embedded visualization because it naturally embeds information within a large-scale indoor environment. In the tasks, participants needed to select one of four targets distributed across a two-floor building based on multiple criteria under time pressure. An AI with 75% accuracy was provided to suggest the optimal target, and both visualizations were experienced using an Apple Vision Pro (AVP) [1], a state-of-the-art AR headset. This abstract task models time-pressured spatial choices with imperfect AI support, where over-reliance can hide errors and under-reliance can waste useful guidance.

Contrary to our expectations, participants exhibited greater inappropriate reliance on AI in the X-ray condition, suggesting that new perceptual and cognitive challenges arise when digital cues are spatially embedded in large-scale physical environments. Further analysis of quantitative data (decision accuracy, response times, reliance metrics, and self-reported confidence) reveals that inappropriate reliance manifested primarily as over-reliance, which means, blindly accepting AI suggestions. Through qualitative interviews, we identify several key factors that contributed to the over-reliance, such as perceptual challenges in AR, visual proximity illusions, and heightened trust in "embodied" AI suggestions. Nonetheless, our findings suggest the X-ray's strength in spatial mapping (as evidenced by quantitative data) and support for egocentric spatial imagery (as reported by participants). These findings matter for AR+AI design because they show that while embedding AI cues into the physical world can strengthen spatial mapping, it can also simultaneously miscalibrate reliance. We conclude by outlining the lessons learned, design implications, and promising future directions for better harnessing these AR strengths to foster human–AI collaboration in spatial decision-making.

2 Related Work

2.1 Human-AI Decision-Making

Human-AI decision-making refers to collaborative decision processes where a human decision-maker interacts with an AI system to make a choice [3]. A common goal in these systems is to achieve *complementary performance*, where the human-AI team performs better than either the human or the AI alone [3]. However, empirical studies have repeatedly found that such complementary performance is rarely observed [5, 12]. One widely reported barrier is *inappropriate reliance* — the inability of human users to correctly accept AI suggestions when they are right, or reject them when they are wrong [3].

Unlike *trust* in AI, which reflects users' subjective feedback, *reliance* objectively measures whether a participant accepts or rejects the AI's suggestion [3, 59], and has therefore been widely studied. To foster *appropriate reliance*, researchers have proposed providing explanations of the AI's suggestions [77]. Unfortunately, several studies found that explanations have often failed to deliver their intended benefits, but actually increase over-reliance – a phenomenon where users uncritically accept AI suggestions, even when they are suboptimal [3]. Drawn on the dual-process model of cognition [13, 24], initial efforts attributed this over-reliance to a lack of cognitive engagement with the explanation, because people often rely on heuristics (Type 1) to judge the AI's suggestions, rather than engaging in deeper analysis (Type 2). Buçinca *et al.* [10] introduced cognitive forcing functions that encourage users to think more critically about the AI's suggestions. Vasconcelos *et al.* [89] proposed a cost-benefit framework that formalizes when users choose to engage with AI explanations. Recently, Guo *et al.* [30] developed a theoretic framework to measure AI reliance by separating reliance behavior from cognitive limitations in signal interpretation.

In the context of visualization research, automated methods to assist humans in making decisions with data have long been practiced. While a large body of work has focused on developing visual

[1]https://www.apple.com/apple-vision-pro/

analytics systems [75], relatively fewer studies have investigated how visualizations affect human-AI collaboration. Among these studies, Yang *et al.* [103] explored how visual explanations can foster appropriate trust in AI systems; Morrison *et al.* [61] investigated how different human explanation strategies might impact users' reliance on AI in visual decision-making; Gaba *et al.* [26], Wang *et al.* [93], and Wall *et al.* [91] each focused on how interface design influences users' perceptions of model bias and performance, highlighting the powerful role of visual presentation in shaping trust in AI systems. Recently, Ha *et al.* [31] found that in AI-guided visual analytics systems, users were more inclined to accept suggestions when completing a more difficult task despite the AI's lower suggestion accuracy. Zhao *et al.* [107] and Reyes *et al.* [73] both investigated how visualizing uncertainty in model outputs influences reliance and trust. Taken together, these studies underscore that visual presentation profoundly influences how users evaluate AI suggestions, rely on them, and ultimately make decisions.

Unlike all these existing works, which evaluate desktop-based human-AI collaboration, our study contributes an initial empirical investigation of the impact of AR visualizations on human-AI collaboration in spatial decision-making tasks. The differences extend beyond display technique (2D monitor vs. 3D AR) to the task environment (desktop vs. physical environment) and to cognitive processes (non-spatial vs. spatial cognition), opening new questions that merit dedicated study.

2.2 Visualization-Aided Decision-Making

In the field of visualization, research on decision-making typically falls into two categories: understanding how humans use visualizations to make decisions [64] and design interactive systems [63].

The former body of work has drawn from theories in cognitive psychology to model and explain the mechanisms by which visualizations influence decision-making. For example, Padilla *et al.* [64] proposed a cognitive framework grounded in dual-process theories, illustrating how visualizations engage both intuitive (Type 1) and analytical (Type 2) reasoning processes in different contexts. Bancilhon *et al.* [2] expanded on this perspective by advocating for the integration of cognitive models in evaluating visualization effectiveness. A parallel line of work investigates how visualizations interact with human cognitive biases. Dimara *et al.* [17] introduced a task-based taxonomy of cognitive biases relevant to visualization, while Wall *et al.* [92] explored how interaction traces might mitigate such biases. Bearfield *et al.* [4] further demonstrated that the same data visualization can elicit diverging interpretations. Research on uncertainty representations also plays an important role in understanding decision-making: Kale *et al.* [41] showed that design choices, such as emphasizing means, can bias effect size judgments, while Fernandes *et al.* [25] found that quantile dotplots can significantly improve real-world decisions in uncertain scenarios.

The latter body of work focuses on developing interactive visualization tools to support decision-making. Some of these systems are domain-agnostic and designed to facilitate specific decision processes, such as inspecting rankings [90] and comparing alternatives based on multiple, often conflicting attributes [66] in Multi-Criteria Decision-Making. Another line of work targets domain-specific decision contexts, such as medical treatment planning [46] , urban policy [95] , and more [14]. These systems typically incorporate domain-specific constraints to guide and contextualize the decision process. A recent survey by Oral *et al.* [63] offers a comprehensive analysis of decision-focused visualization tools and highlights the opportunities and challenges in supporting all stages of the decision-making process.

While decision-making is widely recognized as a core goal of visualization, Dimara and Stasko [18] argue that it remains underrepresented in visualization research and task taxonomies. Brumar *et al.* [9] recently proposed a typology to better describe decision-making tasks in visualization. Despite this growing interest, decision support for spatial tasks remains an underexplored yet promising area [28]. Our study takes a first step toward investigating spatial decision support, aiming to provide a baseline and reference point for future research.

2.3 Spatial Cognition and Spatial Decision-Making

Spatial cognition encompasses the mental processes involved in perceiving, encoding, and reasoning about spatial environments [88]. A fundamental aspect of spatial cognition is the use of reference frames—mental models that help individuals locate and orient themselves in space [87]. Two commonly described reference frames are *egocentric*, where objects are located relative to the observer's current position and orientation like "to my left or right", and *allocentric*, where the environment is represented independently of the observer's position, such as map-like or object-to-object relationships [44]. People switch flexibly between these frames, but doing so can introduce substantial cognitive load [87], particularly when tasks require precise mental transformations, such as rotating or scaling spatial representations in mind.

Spatial decision-making requires integrating information about physical environments with external data. To make such decisions with AI support, decision-makers must reconcile digital information with real-world contexts. With traditional 2D map-based visualizations, this process often involves switching between reference frames, which is cognitively demanding and can affect interpreting or verifying AI suggestions. By contrast, AR can embed information directly in the physical environment, thereby eliminating the spatial mapping overhead [6]. Thus, there is a sustained interest in leveraging AR systems to support spatial decision-making tasks.

2.4 AR, Situated, and Embedded Visualizations

Different from Virtual Reality (VR) [11], AR can directly visualize information within physical context, benefiting a broad range of applications that require computational support [42]. The visualization community has explored AR visualizations across multiple dimensions, including design patterns [50], techniques [108], interactions [56, 110], empirical evaluations [68, 70], and applications [51, 55]. We refer readers to the recent survey [79] for more details.

More broadly, the concept of visualizing information in its physical context is known as *Situated Visualization* [8], for which AR is one prominent realization among others like physicalization [39]. Depending on the spatial relationship between the visualization and its physical referent, Willett *et al.* [97] distinguish between *situated* (co-located in physical space) and *embedded* (overlaid or integrated with the referent). Based on this definition, our study compares two AR visualizations: a *situated* 2D Minimap and an *embedded* AR X-ray view. This comparison enables us to investigate whether embedding data directly into the physical referent can foster more appropriate reliance on AI by reducing the mental effort required to map data from a separate panel onto the physical environment. For clarity in the remainder of the paper, **we refer to these two AR visualizations simply as *Minimap* and *X-ray***, respectively.

2.5 AR Visualizations for Spatial Decision-Making

Our particular interest lies in leveraging AR to support indoor spatial decision-making under time pressure, a crucial area with multiple applications such as emergency evacuation [78, 99], first response operations [106], and security management [84]. Despite its importance, research specifically addressing time-constraint spatial decision-making in AR remains limited. To the best of our knowledge, the most relevant work involves AR-based evacuation assistance, which provides users with dynamic, context-sensitive navigation cues to safely guide them toward exits during emergencies [106]. For instance, Wächter *et al.* [99] investigated AR-guided evacuation in buildings, showing that adapting AR visualizations to environmental hazards can improve evacuation efficiency and safety. Similarly, Sharma *et al.* [78] developed AR modules to enhance situational awareness and cognitive mapping during evacuations in buildings.

Beyond emergency scenarios, a broadly related line of work explores AR-based indoor navigation, where users typically follow predefined routes rather than making spatial decisions. Most AR navigation techniques fall into two categories: annotation-based methods like arrows, and embedded visualizations. Annotations are arguably the most common AR navigation technique, as seen in applications like Google Maps[2]. Numerous studies [19, 52] have compared annotation-based AR navigation to traditional 2D Minimap across various contexts. Nonetheless, AR embedded visualizations offer benefits like a global view but have received less attention, partly due to technical challenges requiring the development of digital twins of real-world environments. Representative work in this direction includes Xu *et al.* [100, 101] who conducted comparative studies of AR X-ray views (a type of AR embedded visualizations) for indoor wayfinding, demonstrating that egocentric perspectives significantly enhanced navigation efficiency, reduced cognitive load, and improved spatial awareness.

Drawing inspiration from these works, we adopt AR X-ray views to present spatial information within indoor environments. Unlike previous studies, our research specifically investigates how this AR embedded visualization can support users in making spatial decisions with AI assistance – a topic that has not yet been explored in depth. Our study thus aims to provide an initial reference point for future research into AI-aided decision-making in AR contexts.

3 Study Rationale

This section defines the scope and abstraction of the spatial decision-making task, and summarizes the rationale behind our choices of visualizations, decision supports, and study hypotheses.

[2]https://maps.google.com

3.1 Task Abstraction and Definition

Spatial decision-making has long been studied under the broader umbrella of *wayfinding* in spatial cognition [23, 96]. Foundational work by Siegel and White [80] conceptualizes wayfinding as a decision process involving three forms of spatial knowledge: landmark, route, and survey. Motivated by real-world applications, our study particularly focuses on **target selection** (landmark). In tasks such as exploring a building, security surveillance, and emergency response, spatial decision-making often involves identifying or selecting a target before proceeding with an action like navigation or intervention. Although AR systems have frequently been investigated for supporting the action phase such as route guidance [100], their potential to enhance the preceding decision-making step remains underexplored.

To address this gap, we abstract real-world scenarios into a domain-agnostic spatial target selection task, defined by the following core attributes derived from the literature:

- **Space - Large-scale Indoor Environment**: These tasks often take place in physical spaces larger than the immediate space around a person such as buildings or open areas [27, 29]. Tversky [87] characterize such spaces as *space of navigation.* In this study, we focus on indoor building environments, given their prevalence across numerous application domains.
- **Time - Sensitive**: Many high-stakes spatial tasks impose stringent time constraints, requiring users to make optimal decision as soon as possible [62]. In this study, we focus on time-pressure scenarios that particularly benefit from decision support systems.
- **Data - Static and Dynamic**: Appropriate target selection relies on both static data like building level layouts, and dynamic updates from sensors like walking people, occupant counts, and hazard alerts to evaluate each target [1].
- **Decision - Multi-Criteria**: These tasks frequently require decisions among multiple targets or locations distributed over rooms, floors, or areas [48]. Success often involves balancing or prioritizing several factors, such as response time, travel distance, and urgency [67].

By focusing on these domain-agnostic factors, we can investigate how AR-based decision support influences spatial decision-making without limiting our findings to one specific use case.

3.2 Visualization Methods: Minimap vs. X-Ray

Given that spatial decision-making inherently involves interpreting spatial relationships, an effective decision support tool must present relevant data within its spatial context. Thus, we select map-based visualizations as the foundational interface for our study. We are particularly interested in how **embedded visualizations** influence users' decision-making processes. Therefore, we choose two representative visualization methods: a Minimap and a X-ray. Below, we discuss the rationale behind choosing these two methods and briefly highlight alternative visualization approaches.

Minimap. 2D Minimap is perhaps the most widely adopted map-based visualization on traditional flat-screens. It presents the spatial environment as a simplified, abstracted, top-down representation, upon which additional data can be annotated, such as Google Maps. For indoor environments spanning multiple floors, 2D Minimaps often employ separate layers for each floor, indicating spatial relationships through key landmarks such as stairways or elevators. The use of 2D Minimap is well-established and requires minimal training, making them a suitable baseline for comparison. In our study, we implement a Minimap (Figure 1 Right) that shows the spatial distribution of multiple targets and their associated data using glyphs.

X-Ray. Given our task focuses on large-scale indoor environments, we select AR X-ray (Figure 1 Left) as our embedded visualizations, which overlays a 1:1 scale 3D map on the real-world. It allows users see the interior layouts or objects behind walls, as if they had "X-ray" vision [57], which has been widely used in applications like navigation [101] and construction [82]. For our study, we follow previous similar research [101] and portray the building's indoor layout, along with multiple targets (represented as cubes), and their associated data (the number of people in line) using virtual avatars.

Alternatives. World-in-Miniature (WiM) [15, 83] provides a three-dimensional representation of the environment and can be implemented on flat-screen or AR displays. However, WiM typically relies heavily on interactions such as panning, tilting, zooming, or rotation to fully leverage their advantages. Because our current investigation prioritizes evaluating the impact of the visual representations themselves (without the influence of interactive complexity), we intentionally avoided such interactions in our study. Thus, we reserve exploration of interactive WiM and other advanced visualization methods for future research.

3.3 Visual Decision Supports

A recent survey by Oral *et al.* [63] highlights a broad spectrum of visual decision support tools, ranging from basic information visualizations to advanced AI-based suggestions [49]. For this study, we specifically choose to implement two fundamental yet representative forms of decision support across both the Minimap and X-ray:

- **Target Visualizations** that present candidate targets and associated real-time data like occupancy counts on the map, without path suggestions or explanatory text.
- **AI-Suggestions** that highlight a single optimal target based on predefined criteria, with no accompanying explanation. Similar to all AI systems, suggestions may not always be accurate. We detail our adjustment of the performance of the AI based on prior works [85] and pilot studies in Sec. 4.7.

Following practices established in prior research [89], we deliberately constrain our decision support to these fundamental features for several reasons: First, target and real-time data visualization represent core functionalities broadly used in existing decision-support tools. Second, simple AI suggestions enable direct investigation of user reliance on AI assistance under different visualizations, without the additional cognitive complexity and potential confounding effects introduced by detailed explanations [40].

While more advanced support, such as comparative visualizations, AI-generated explanations, and uncertainty displays, may offer added benefits, we intentionally focus on these fundamental forms to isolate the role of embedded visualizations in spatial decision-making and AI reliance. This controlled setup provides a baseline for future studies exploring richer decision supports.

4 Study Design

We conducted a 2 × 2 within-subjects study to investigate whether X-ray can foster more appropriate reliance on AI in spatial decision-making. The two independent variables were the type of visualizations (`X-ray` vs. `Minimap`), and the availability of AI assistance (`NoAI` vs. `AI`). The study design is detailed below, with fixed values based on a 13-participant pilot study. The study was approved by university ethics review board (IRB).

4.1 Task Description

We designed a realistic indoor building scenario, where participants acted as busy employees. The goal was to simulate a time-pressured spatial decision-making scenario that closely resembles everyday scenarios to the participants. In each trial, participants were tasked with **selecting one coffee machine from four available options within a two-floor building within 20 seconds**. The objective was to choose the machine that would allow them to obtain a cup of coffee in the shortest possible time. Optimal decision-making in this task depended on two key factors:

- **Walking Distance** — the length of the walking path from the participant's current location towards the spatial location of each coffee machine.
- **Queue Length** — the number of people currently waiting in line at each machine, which thus decides waiting time.

Participants accessed information about both the machine locations and queue lengths via one of two visualizations, and in half of the trials, they also received AI suggestions.

4.2 Spatial Arrangement

The task was set in a real two-floor indoor building spanning roughly 80m × 80m. In each trial, the participant and the four coffee machine targets were positioned based on a predefined *spatial arrangement.*

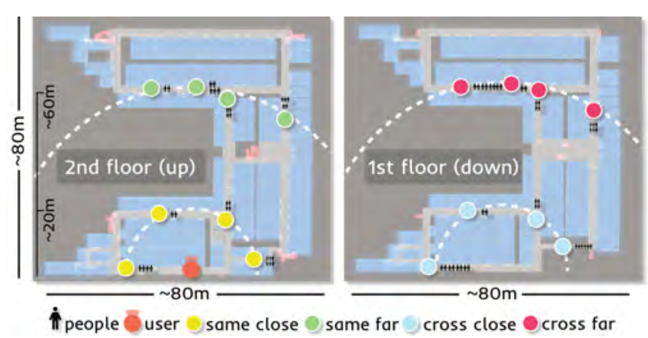

Figure 2: An example spatial arrangement of the coffee machine targets locations for a single participant location, illustrating four distinct difficulty levels: SAME+CLOSE, SAME+FAR, CROSS+CLOSE, and CROSS+FAR. For each participant location, there are 4 trials, one difficulty level per trial. The participant's location is marked in red, and colored dots indicate target locations in each difficulty level. The white dashed lines indicate that the coffee machine targets are placed at equal line-of-sight distances from the participant.

Target and Participant Locations. To simulate diverse spatial scenarios and mitigate learning effects, we varied the distribution of the four coffee machines across trials based on two spatial factors identified during the pilot testing:

- **Floor Relationship – SAME or CROSS:** All four targets were located on the same floor, which was either the same as the participant's floor (SAME) or a different one (CROSS). We anticipated that this factor would substantially impact the difficulty of estimating walking distances, particularly in the `Minimap` condition.
- **Proximity – CLOSE or FAR:** Targets were categorized as either CLOSE or FAR in direct proximity (instead of walking distance) to the participant's location. We anticipated that this factor would have a greater impact on task difficulty in the `X-ray` condition, as distant targets would subtend a smaller retinal angle than closer targets and thus be harder to perceive.

The combination of these two factors resulted in four difficulty levels, ranging from SAME+CLOSE (easiest) to CROSS+FAR (most difficult). Within each trial, to ensure that the appearance of each target was perceptually similar to the user, we used the same difficulty level setting to define each target's location and placed them all at equal line-of-sight distances from the participant's location. The participant's location was then positioned at the center of a circle passing through all four targets. Figure 2 illustrates an example of the four difficulty levels.

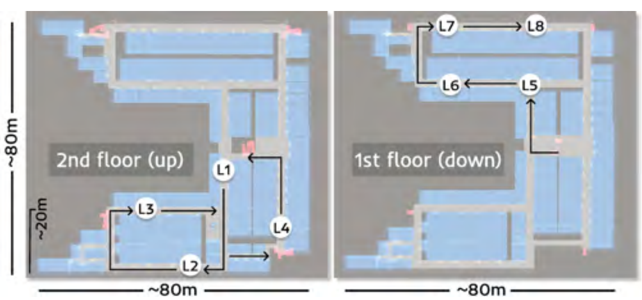

Figure 3: The 8 participant locations in the study, distributed across two floors of the building. Participants were instructed to complete 4 trials at each location, following the order indicated by the numbered labels and route. The letter "L" is an abbreviation for location.

Grouping Arrangements by Participant Location. Ideally, both participant and target locations would vary each trial. However, moving participants between locations would introduce substantial overhead and fatigue. To address this, we grouped the target arrangements by participant location. Specifically, we selected eight participant locations evenly distributed in the building (Figure 3), and for each location, we designed four spatial arrangements of targets, each representing one difficulty level. This resulted in $4 \times 8 = 32$ unique spatial arrangements. An example of one group is shown in Figure 2.

Target Optimality. Target optimality was determined by two metrics: walking distance from the participant and queue length. For simplification, we estimate the total time to obtain coffee from a target using this formula:

$$T_{\text{coffee}}(d, n) = \frac{d}{\text{SPEED}} + n \times \text{SERVICE\_TIME}$$

where d denotes the walking distance to the target, and n means the number of people in line. SPEED and SERVICE_TIME are constants denote the average walking speed and the average waiting time per person in line, respectively. Based on our pilot studies, we empirically set SPEED at 1m/s and SERVICE_TIME as 15s. While individuals naturally differ in walking speed, participants in our study did not need to physically walk; SPEED was used for estimation and is thus not expected to introduce substantial bias. These parameters were explicitly communicated to participants in the study.

Since the values of d were fixed by the spatial arrangement of the targets, we manipulated the queue lengths n to differentiate target optimality. Following prior research [85], each trial was designed to include intentionally varied options: **one optimal, two suboptimal, and one worst choice**, differing by at least 10 seconds in T_{coffee}. This setup allowed us to evaluate potential participant over-reliance on AI suggestions. Detailed calculations are provided in the supplementary materials.

4.3 Implementation and Apparatus

Following similar work [10], a simulated AI was used to provide decision suggestions in applicable trials. Details of the simulated AI setup are provided in Sec. 4.7. We developed the minimap based on high-precision CAD models provided by the building manager. The AR visualizations was developed using ARKit [3] and AR Mesh Manager [4] to scan the building and construct a 3D digital twin in Unity 6000.0.29f1, with ARAnchor Manager [5] ensuring proper registration of the digital model with the physical environment. The system is available at our public repository https://github.com/demoPlz/ARAIReliance. All trials were completed using an Apple Vision Pro to eliminate device-related differences and provided a high-fidelity immersive experience.

4.4 Participants

We recruited 32 participants (16 male, 16 female; ages 18–28) from our university community via email, word of mouth, and posted flyers. For base compensation, participants received a $15 Amazon gift card. To incentivize accurate performance, participants earned an additional $0.20 for each trial in which they selected the optimal target, with the bonus capped at $20.

To use Apple Vision Pro, participants were required to have normal vision, use contact lenses, or use the prescription inserts provided by us. To minimize potential bias arising from participants' familiarity with the experimental setting, we exclusively recruited individuals whose prior exposure to the building used in the study was limited. Specifically, 19 participants (9 male, 10 female) had never visited the building, 12 participants (6 male, 6 female) had

[3] https://developer.apple.com/augmented-reality/arkit/
[4] https://docs.unity3d.com/Packages/com.unity.xr.arkit@6.2/manual/arkit-meshing.html
[5] https://docs.unity3d.com/Packages/com.unity.xr.arfoundation@6.2/manual/features/anchors/aranchormanager.html

Figure 4: A stimulus workflow from an actual trial. Each trial began with a 20-second countdown. Participants viewed either the `X-ray` or `Minimap` visualization while searching for the optimal target. After identifying their choice, participants verbally announced the selected target and rated their confidence on a 1–7 scale. Finally, they physically pointed to the perceived location of the chosen target. For illustration, the figure shows only one trial at location L, but participants completed four trials at each location.

been in the building between one and five times, and 1 participant (male) had visited the building between five and ten times.

4.5 Procedures

With a within-subjects design, each participant completed trials under four conditions: **`Minimap+NoAI`**, **`Minimap+AI`**, **`X-ray+NoAI`**, and **`X-ray+AI`**. The order of conditions was counterbalanced across participants using a Latin square design to minimize order effects and learning biases. Our primary focus was on participants' behavior in the `AI` conditions, while the `NoAI` conditions served as baselines to better understand user decision-making patterns.

We used the following full-factorial within-subject study design with Latin square-randomized order of the conditions to minimize learning and fatigue effects:

	32	Participants
×	4	Conditions: { `Minimap+NoAI`, `Minimap+AI`, `X-ray+NoAI`, `X-ray+AI` }
×	2	Participant Location
×	4	Spatial Arrangements: { Same+Close, Same+Far, Cross+Close, Cross+Far }
	1024	total trials (32 per participant)

Participants were split evenly between the 4 Latin square-randomized condition orders. The orders of the spatial arrangements and participant locations were kept consistent across all participants to control for any confounding effects related to environmental layout (see Figure 5).

Procedure Overview. Prior to the experiment, participants were required to take an elevator ride from the ground floor to the experiment floor (2nd floor) and wait at the elevator door for the research team. This procedure was designed to minimize any prior exploration or accumulation of spatial knowledge about the experimental environment. The study lasted approximately one hour in total:

- *Introduction (10mins)*: Participants first provided informed consent and were explicitly informed about the performance-based incentives. Next, participants received a detailed tutorial that explained the AVP's calibration, its usage, and task procedures.
- *Training Tasks (10mins)*: Participants completed four training trials, one per condition, to familiarize themselves with the task and visualizations. The participant locations and target locations were different in the training trials than in the experimental trials (see supplementary materials).

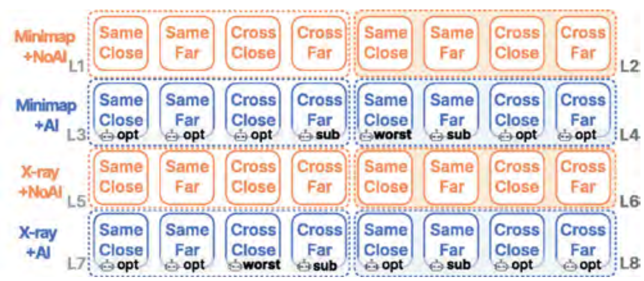

Figure 5: One example of what a participant experienced during the experiment. Between participants, the order of the 4 conditions (`Minimap+NoAI`, `Minimap+AI`, `X-ray+NoAI`, `X-ray+AI`) was varied and counterbalanced using a Latin square design. Factors that remained fixed across all participants included the sequence of standing locations (L1-L8) and the order of the four spatial arrangements (Same+Close, Same+Far, Cross+Close, Cross+Far) at each standing location. Within each AI condition, the presentation order of the AI suggestion optimality levels (opt, sub, worst) was randomized, while ensuring a balanced distribution across participants. For illustration, only two examples of AI suggestion presentation orders are shown.

In these trials, after participants made their selections, the optimal target was explicitly highlighted to illustrate optimal decision-making. Participants were encouraged to ask questions and were informed they could withdraw at any point.

- *Actual Tasks (30mins)*: Participants completed 32 trials, grouped into four blocks, with each block distributed over 2 consecutive locations (see the example in Figure 5). Within each block, one condition was tested, namely, eight trials per condition. Within each location, they remained physically stationary and completed four trials (one trial per difficulty level). Breaks were allowed during the study.
- *Post-Study Interview (10mins)*: Participants were interviewed with a list of open-ended questions (attached in supplementary materials).

Detailed Trial Procedure. As illustrated in Figure 4, each trial proceeded through the following steps: The experimenter escorted the participant to the designated starting point and helped them put on the headset. When ready, the experimenter used a controller to initiate the trial, activating the visualization and a 20-second countdown timer. Participants verbally announced their selection, for example, *"Target A"*, as soon as they identified the optimal target. Then, the experimenter stopped the timer, automatically clearing the visualization. Participants verbally rated their confidence in their choice using a 1 (Strongly unconfident) to 7 (Strongly confident) scale. Lastly, participants physically pointed toward the actual spatial location of their chosen target and looked at their pointing direction. This process was repeated for each trial. We recorded the whole session through audio and first-person view video with participants' informed consent.

4.6 Measures

For every trial, we documented and calculated the following measures based on prior research [76, 85]:

- **Decision Accuracy**: We defined decision accuracy as the *average point score* across trials in each condition, ranging from 0 to 1. Following previous research [85], participants earned points according to their choice quality: optimal choice (1 point), suboptimal choices (0.5 points), and the worst choice (0 points). Each trial contained exactly one optimal target, two suboptimal targets, and one worst target.
- **AI Reliance**: Following previous studies [103], we categorized participant reliance on AI as either appropriate or inappropriate (Figure 6). We classified reliance as appropriate when participants either followed correct AI suggestions or overrode incorrect ones with better choices. Inappropriate reliance included: (1) over-reliance: accepting suboptimal or worst suggestions; and (2) under-reliance: rejecting correct suggestions in favor of a worse option.
- **Response Time**: We measured the duration from the start of each trial to the moment the participant verbally announced their choice.
- **Confidence**: Participants rated their confidence in their decision on a 7-point scale from 1 (Strongly unconfident) to 7 (Strongly confident).
- **Point Error**: Similar to prior research [36, 94], we assessed whether participants correctly pointed to the spatial location of their selected target based on first-person video recordings. Two co-authors independently coded each response as 0 (correct) or 1 (incorrect) to ensure inter-rater reliability. Inter-rater reliability was evaluated via Cohen's Kappa ($\kappa = 0.898$), suggesting almost perfect agreement according to Landis and Koch's interpretation [47]. Discrepancies were discussed and resolved through consensus. This measure indicates if participants actually knew the spatial location of their selected target.

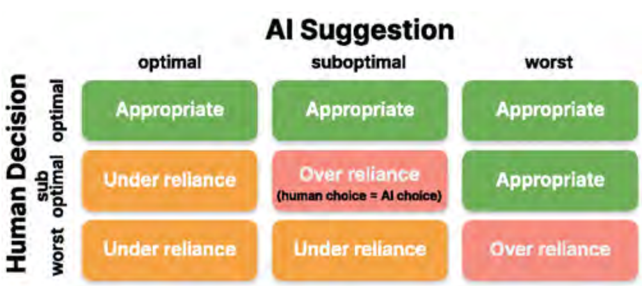

Figure 6: Classification of reliance as appropriate, over-, and under-reliance based on the alignment between AI suggestions and human decisions. Appropriate reliance on AI was defined as participants accepting optimal suggestions or overriding suboptimal and worst suggestions. Inappropriate reliance included: (1) over-reliance: accepting suboptimal or worst suggestions; and (2) under-reliance: rejecting AI suggestions in favor of a worse option.

4.7 Using Simulated AI

Similar to prior works [10, 85], we used a simulated AI with an average accuracy of 0.75 to ensure controlled experimentation. In each AI condition, the AI suggested the optimal target in five trials, a suboptimal target in two trials, and the worst target in one trial (Figure 5). This setup yields an average accuracy of $0.75 = \frac{5\times1+2\times0.5+1\times0}{8}$. The order of these suggestions was randomized and counterbalanced across participants, ensuring that optimal, suboptimal, and worst suggestions were evenly distributed across all trials and participants.

Mitigating Effects of Trust in AI. Following previous work [3, 10], we took steps to neutralize trust effects so that observed differences in reliance could be attributed primarily to the visualizations. Specifically, we avoided anthropomorphizing the AI [35], consistently referring to it as "the AI". We also described AI outputs as "suggestions" to counteract perfect automation schemas [21] and emphasize that these suggestions could be incorrect. Furthermore, we do not provide decision feedback for each trial [16], or any information about the AI's accuracy [105].

4.8 Hypothesis

According to dual-process theory, human decision-making can be categorized into two distinct cognitive systems: rapid, heuristic-driven *Type 1* thinking, and slower, analytical *Type 2* thinking [13]. A widely accepted distinction between these systems is that Type 2 thinking demands significantly more cognitive resources than Type 1 [13].

Under conditions of high time pressure, users often lack sufficient cognitive bandwidth to analytically evaluate AI-generated suggestions, potentially resulting in inappropriate reliance, either rejecting correct suggestions or accepting suboptimal ones [58]. This risk is especially pronounced with 2D minimaps, as these interfaces require users to mentally map the information to their corresponding physical environment, imposing substantial cognitive load due to reference-frame switching [32]. In contrast, embedded visualizations can substantially reduce or eliminate these spatial mapping demands by directly fusing information within the physical environment, thus potentially fostering more *appropriate reliance* on AI suggestions.

Based on this rationale, we hypothesize that in AI-aided spatial decision-making, compared to the 2D minimap:

- **H1** AR X-ray improves overall task performance, specifically decision accuracy.
- **H2** AR X-ray promotes more appropriate reliance on AI.
- **H3** AR X-ray shortens response times in decision-making.

5 Results

5.1 Quantitative Results

Prior to analysis, we excluded trials in which participants failed to make a decision within 20 seconds. After filtering, we retained 248 trials in `Minimap+AI`, 247 in `Minimap+NoAI`, 237 in `X-ray+AI`, and 237 in `X-ray+NoAI`, out of 1024 (32 participants × 32 trials). For each measure, trials were aggregated by participant and conditions being analyzed. We first tested for normality using the Shapiro-Wilk test. Results indicated that accuracy, point error, and AI reliance significantly violated the assumption of normality, whereas time and confidence did not. Accordingly, we used repeated-measures ANOVA ($\alpha = .05$) to assess the significance effects on time and confidence; then, post-hoc pairwise comparisons were conducted using Holm-Bonferroni corrections. For accuracy and point error, we used Aligned Rank Transform (ART) ANOVA [98] to test significance and use ART-C [22] for post-hoc contrast testing. To compare AI reliance in the `Minimap+AI` and `X-ray+AI` conditions, we used Wilcoxon signed-rank tests.

5.1.1 Decision Accuracy. We computed the mean accuracy of target selection for each condition at the participant level by averaging across all

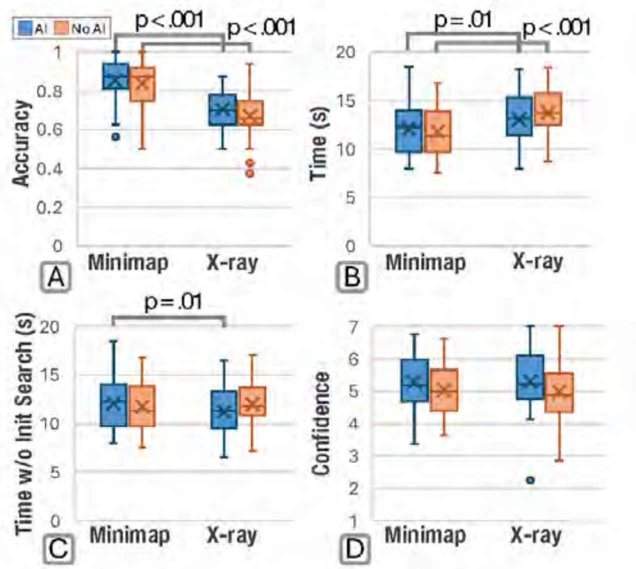

Figure 7: Summary of quantitative results across conditions. Metrics include (A) decision accuracy, (B) task completion time, (C) completion time excluding initial search, and (D) confidence ratings. Comparisons are shown between AI-assisted and non-AI conditions for both `Minimap` and `X-ray` visualizations, with significance levels indicated. The "X" denotes the mean, the horizontal line inside each box indicates the median, the box spans the interquartile range (IQR). Outliers are defined as values beyond the upper cutoff $Q_3 + 1.5 \times \mathrm{IQR}$ **or the lower cutoff** $Q_1 - 1.5 \times \mathrm{IQR}$**.**

valid trials for each participant (Figure 7a). The analysis showed a significant effect between *Visualizations* on the decision accuracy ($F(1,93) = 82.07, p < .001$). Post-hoc analysis further revealed that participants were more accurate in the `Minimap+AI` condition ($Mdn = 0.88, IQR = 0.13$) than in `X-ray+AI` ($Mdn = 0.70, IQR = 0.13$), with a significant effect ($F(1,93) = 6.26, p < .001$). Participants were more accurate in `Minimap+NoAI` ($Mdn = 0.88, IQR = 0.16$) than in `X-ray+NoAI` ($Mdn = 0.66, IQR = 0.13$), with a significant effect ($F(1,93) = 6.56, p < .001$). These results **reject our first hypothesis H1**.

Further observation suggests that **`X-ray+AI` under-performance may stem from inappropriate reliance on AI**. Participants in the `X-ray+NoAI` condition achieved a median decision accuracy of $Mdn = 0.66$, which increased to $Mdn = 0.70$ in the `X-ray+AI` condition. However, this improvement still *fell short of* the AI standalone baseline accuracy of 0.75, suggesting that participants may have relied on it inappropriately. This aligns with findings in prior empirical studies [3]. In contrast, performance in both `Minimap+NoAI` ($Mdn = 0.88$) and `Minimap+AI` ($Mdn = 0.88$) conditions exceeded the AI baseline, indicating that participants were able to appropriately rely on AI by identifying and overriding poor suggestions.

To further examine the nature of participants' reliance, specifically whether it reflected under- or over-reliance, we broke down trials by AI correctness into AI-correct, no-AI, and AI-wrong. The analysis indicated there is a significant effect of *AI Correctness* ($F(2,155) = 47.97, p < .001$), *Visualizations* ($F(1,155) = 73.36, p < .001$), and *AI Correctness*× *Visualizations* interaction ($F(2,155) = 3.43, p = 0.03$) on decision accuracy. Post-hoc analysis further revealed that there are significant differences in the `X-ray+AI` condition. Compared to the `X-ray+NoAI` baseline, participants treated AI input as authoritative: they performed better when the AI was correct ($\Delta_{Mdn} = +0.20, F(1,155) = 4.76, p < .001$), but worse when the AI was wrong ($\Delta_{Mdn} = -0.16, F(1,155) = -3.37, p = .01$). This result indicated that participants tended to defer their decisions to the AI, even when the AI is misleading. In contrast, for `Minimap`, we found no significant differences between AI-correct and no-AI conditions, nor between AI-wrong and no-AI conditions. In sum, the breakdown analysis indicated that **participants exhibited greater inappropriate reliance on AI in `X-ray+AI` compared to `Minimap+AI`**. Next, we further analyzed reliance patterns using dedicated reliance metrics.

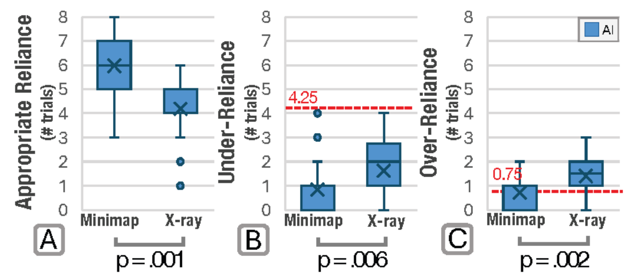

Figure 8: Number of trials per block in which participants exhibited each type of AI reliance, including (A) appropriate reliance, (B) under-reliance, and (C) over-reliance. Random-selection baselines (in red) are indicated. The "X" denotes the mean, the horizontal line inside each box indicates the median, the box spans the interquartile range (IQR). Outliers are defined as values beyond the upper cutoff $Q_3 + 1.5 \times \mathrm{IQR}$ **or the lower cutoff** $Q_1 - 1.5 \times \mathrm{IQR}$**.**

5.1.2 AI Reliance. Results from the accuracy analysis indicated that participants relied on AI inappropriately in the `X-ray+AI` condition. Quantitative analysis on the AI Reliance measure confirms this pattern: we observed a significant difference between `Minimap+AI` and `X-ray+AI`, with participants showing less appropriate reliance on AI in `X-ray+AI` (Figure 8a). This finding again **contradicts our second hypothesis H2**.

Inappropriate reliance on AI can be either over-reliance or under-reliance. To examine which pattern dominated participants' interactions with AI in `X-ray+AI`, we separately analyzed trials of over-reliance and under-reliance. Figure 8b and c depict the average occurrences of over- and under-reliance, respectively, per block (8 trials).

To interpret these numbers, we compared the observed behaviors with a random-selection baseline. Among the 8 trials in each block, the AI suggested 5 optimal, 2 sub-optimal, and 1 worst target. If a participant were to choose a target at random, namely with equal probability of 1/4 for each of the four possible targets, the expected numbers of over- and under-reliance trails per block are 0.75 and 4.25 respectively (see Supplemental Materials for calculation details). Compared to this baseline, participants in the `X-ray+AI` condition displayed fewer instances of under-reliance ($M_{\texttt{X-ray}} = 1.63 < M_{\text{random}} = 4.25$), but notably more instances of over-reliance ($M_{\texttt{X-ray}} = 1.41 > M_{\text{random}} = 0.75$). In contrast, participants in the `Minimap+AI` condition exhibited fewer instances of both under-reliance ($M_{\texttt{Minimap}} = 0.84 < M_{\text{random}} = 4.25$) and over-reliance ($M_{\texttt{Minimap}} = 0.72 < M_{\text{random}} = 0.75$). These findings indicate that **inappropriate reliance in the `X-ray+AI` condition was primarily driven by over-reliance**, consistent with our earlier accuracy analysis (Sec. 5.1.1).

5.1.3 Response Time. We computed the average response time across conditions at the participant level, and observed significant effects (Figure 7b). The analysis showed a significant effect between *Visualizations* ($F(1,31) = 44.30, p < .001$) and *Visualizations* × *AI Conditions* interaction ($F(1,31) = 5.92, p = .02$) on the response time. Further post-hoc analysis showed that participants responded faster in `Minimap+AI` (M = 12.08, SE = 0.48) than `X-ray+AI` (M = 13.03, SE = 0.48) with a significant effect (Δ_M = -0.94, $F(1, 31) = -3.12, p = .01$), and also in `Minimap+NoAI` ($M$ = 11.76, SE = 0.44) than `X-ray+NoAI` (M = 13.77, SE = 0.44) with a significant effect (Δ_M = -2.00, $F(1, 31) = -6.28, p < .001$). These results **contradict our third hypothesis H3**.

Response Time Excluding Initial Search. We observed that participants in the `X-ray` conditions spent additional time to visually search the spatially distributed targets—especially those outside their immediate field of view. In contrast, in the `Minimap` conditions, all targets were presented in a 2D overview, making search overhead negligible. Prior decision-making model (Simon's model [81]) and empirical studies [72, 74] suggest that actual *decision-making* begins only after sufficient information has been gathered. Thus, similar to prior research [38], we distinguished the *initial search* phase from the subsequent *decision-making* phase. This motivates us to exclude the initial search overhead in the `X-ray` conditions to better compare its *decision-making* phase with the `Minimap` conditions.

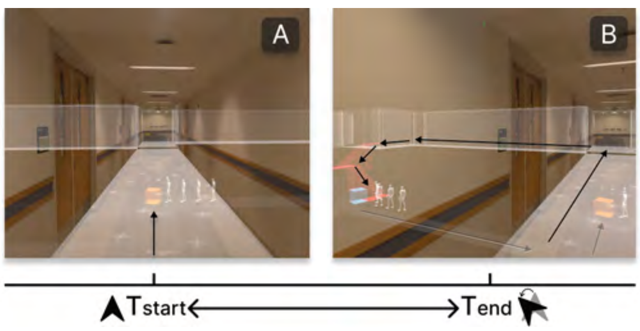

Figure 9: (a) and (b) show the direct and walking distance to the orange target; (a) and (b) also indicate the start and end moment of searching for the blue target.

Following prior visualization-decision work [65], we consider that meaningful decision-making start only after at least two options have been identified. Thus, for each trial in the `X-ray` conditions, we excluded the initial search time required to discover ***the first two distinct targets***. Specifically, we defined two search segments to exclude: S1-from the start of the trial to the first frame when the first target entered the user's field of view; and S2-from the frame when the user turned away from the first target (Figure 9a) to the frame when the second target appeared (Figure 9b). If a target was already visible at trial start, or if the first and second targets appeared simultaneously, the corresponding segment (S1 or S2) was set to zero. This definition provides a strict lower bound on search time, as participants may still require additional head movement to center the target in view. Yet, we acknowledge that this measure is not a theoretically perfect indicator of true decision time. Nevertheless, it provides a scientifically reasonable approximation that is consistent with prior research [38]. We discuss its limitations and possible refinements in Sec. 6.5.

Two co-authors independently annotated these segments frame by frame from the first-person recordings to ensure inter-rater reliability. Inter-rater reliability was assessed using a two-way mixed effects intraclass correlation coefficient, ICC(3,1), which indicated excellent agreement ($ICC(3, 1) = 0.995$, 95% CI [0.992, 0.997]) [45]. Encoding consistency was ensured through frame-level accurate marking and the clear visual cues afforded by first-person video recordings. Discrepancies were discussed and resolved through consensus. After subtracting the annotated initial search time from the response time, we observed a different pattern (Figure 7c). The analysis showed that there is a significant effect of *Visualizations* × *AI Conditions* interaction ($F(1, 31) = 7.70, p = .009$). Further post-hoc analysis indicated that participants in `X-ray+AI` (M = 11.23, SE = 0.48) spent less time than those in `Minimap+AI` (M = 12.08, SE = 0.48) with a significant effect (Δ_M = -0.85, $F(1, 31) = -3.27, p = .016$), indicating that, **with AI assistance, decisions were made more quickly with the X-ray visualization**. No significant difference was observed between `Minimap+NoAI` and `X-ray+NoAI`, suggesting that, without AI assistance, participants spent comparable time making decisions across visualizations. In summary, although raw response times were longer in `X-ray` conditions, a detailed breakdown suggests that this extra time wasn't entirely spent on the decision-making. This leads to interesting implications that we will further discuss in Sec. 6.4.

5.1.4 Post-Trial Pointing Error Rate. We analyzed participants' post-trial pointing error rates across the two visualizations (`Minimap` vs. `X-ray`) at the participant level, regardless of AI availability. We found a significant effect ($F(1, 93) = 40.77, p < .001$), with higher error rates in the `Minimap` condition ($Mdn = 0.08, IQR = 0.20$) compared to the `X-ray` condition ($Mdn = 0, IQR = 0$). This result indicates that **the X-ray visualization significantly improved participants' spatial mapping ability compared to the Minimap**.

5.1.5 Self-Reported Confidence. We also examined participants' self-reported confidence ratings across the four conditions. This measure reflects how confident participants were that they had selected the optimal target in each trial. Our analysis revealed no significant differences across conditions (Figure 7d), suggesting that **participants felt equally confident in their decisions regardless of visualization type or AI assistance**.

5.2 Qualitative Results

We conducted semi-structured post-study interviews to capture participants' strategies, rationales, and motivations behind their decision-making across all conditions. Participants are referred to as P1–P32. We analyzed the interviews using reflexive thematic analysis [7], identifying themes both inductively (bottom-up) and deductively (top-down). Below, we summarize the key themes that emerged from the analysis.

5.2.1 Diverse Decision Strategies Across Visualizations. We observed that participants adopted different decision-making strategies depending on the visualization.

Both Visualizations: People First and Distance First Strategies. Participants primarily adopted one of two strategies: (1) prioritizing fewer people in line, then considering walking distance; or (2) prioritizing shorter distance, then checking queue length. In the `Minimap` conditions, participants were almost evenly split between these two strategies, with roughly half employing a "people first" (15/32) and the other half a "distance first" (14/32). In the `X-ray` condition, more participants adopted a distance-first strategy (15 out of 32), while fewer prioritized queue lengths (11/32). The remaining participants (6/32) reported balancing both factors equally without a dominant criterion.

X-Ray: Unique Heuristics and Motivations. While strategy in the `Minimap` condition appeared to depend largely on participants' spatial skills, the `X-ray` condition introduced distinct perceptual heuristics influencing participants to prioritize distance:

- **Egocentric Spatial Imagery.** The egocentric view in our AR X-ray appeared to support participants in vividly imagining physical navigation, particularly for complex pathways involving hallway turns or floor changes. For example, P9 noted, *"It's easier to imagine how I should walk to the coffee machine, especially when targets are on a different floor from me..."* Such imagination of navigation in the AR X-ray may have led to weigh distance more heavily in their decisions.
- **Visual Proximity Illusions.** Participants reported a misleading proximity effect caused by the X-ray view's "see-through" nature. Because the visualization allows users to see targets behind walls,

some targets appeared directly nearby (Figure 9a) even though the actual walking path to reach them was much longer (Figure 9b). This discrepancy often led users to underestimate the true walking distance, particularly when they cannot easily find the path. As P8 noted, *"I think this way is shorter but it's actually longer."* This difficulty occasionally placed additional cognitive burden on participants and may led them to overweight distance.

- **Occlusion Challenges.** Some participants such as P2, P7 and P11 reported difficulties estimating queue lengths due to occlusion in the `X-ray` conditions, prompting reliance on distance estimates instead of people counts.

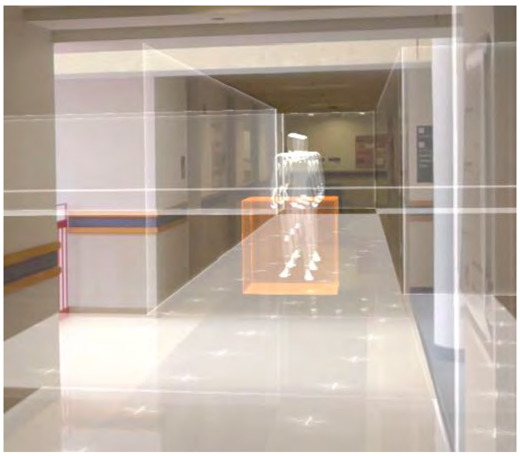

Figure 10: Example of occlusion in the AR X-ray visualization, where overlapping avatars make it difficult to estimate the number of people in the queue, leading to potential misjudgment of wait time.

Summary. Compared to the `Minimap` conditions, participants in the `X-ray` conditions showed a stronger preference for spatial reasoning (distance) in their decision-making. This suggests that the AR X-ray visualization inherently facilitates spatial perception, yet they also function as a double-edged sword by introducing perceptual biases and occlusion challenges.

5.2.2 Increased Trust in AI Suggestions with X-Ray. The visualizations also impacted participants' subjective trust on AI.

`Minimap`: Make Own Choice First, Then Consult AI. In the `Minimap` conditions, most participants (22 out of 32) preferred first to form their own judgment and then cross-check against the AI suggestion, for instance, *"I use it just to double-check my decision"* (P22). Participants reported feeling more confident in their own judgment when using the `Minimap` visualization because it presented the information required to make informed choices. They typically consulted the AI in cases of uncertainty or indecision: *"If I felt stuck between two locations, I would check the AI to break the tie"* (P20).

`X-ray`: AI as a Starting Anchor. Participants expressed greater trust in AI suggestions in the `X-ray` conditions, frequently using the AI suggestion as an initial reference or anchor for their decisions. This subjective feedback may help explain the observed over-reliance. Several participants such as P6 and P11 mentioned that AI helped narrow the decision space and enabled quicker comparisons. Participants provided several reasons for trusting AI more with `X-ray`:

- **Perceptual and Navigation Challenges.** Participants found distance estimation (P2), queue length assessment (P4, P5), and locating navigation paths (P12) more effortful in AR. Thus, AI suggestions offered a useful cognitive shortcut, aligning with findings from prior research on AI-aided decision-making [3].
- **Unfamiliarity with AR X-ray.** Some participants like P31 mentioned that unfamiliarity with the AR X-ray increased their reliance on AI guidance.
- **Enhanced Immersive Realism and Embodiment of AI suggestions.** Interestingly, several participants such as P4 and P13, indicated that the realistic, immersive, embodied presentation of AI suggestions within the AR X-ray heightened their sense of trust and acceptance, which sometimes led to over-reliance on wrong AI suggestions.

Summary. Compared to the `Minimap`, the `X-ray` visualization led to greater participant trust in and reliance on AI suggestions. Participants attributed this shift to greater perceptual complexity, unfamiliarity with the AR X-ray visualization, and the more realistic presentation of AI.

6 Discussion

In this section, we synthesize the key lessons learned, design implications, and promising future directions to advance human-AI collaboration in spatial tasks in AR.

6.1 Perceptual Challenges of AR X-Ray Can Lead to Over-Reliance on AI

Our findings suggest that perceptual challenges of the AR X-ray, such as occlusion, large canvas space, and spatial estimation difficulty, can increase users' reliance on AI suggestions, even when those suggestions are suboptimal. This over-reliance does not appear to stem from blind trust in AI, but rather from the inherent complexity of perceiving and evaluating visual information in large-scale physical environments. While prior research has discussed AR's perceptual limitations [20], our study directly connects these challenges to AI reliance patterns in time-sensitive spatial decision-making. This highlights the need for better embedded visualization designs that support spatial reasoning, especially in large-scale environments.

Design Implications. Visually embedding AI support in physical environments is not inherently beneficial. It can be neutral or even detrimental when perceptual challenges make it harder to verify AI suggestions. These findings underscore the need to rethink how AI assistance is integrated into AR, rather than directly porting AI designs from desktop or 2D interfaces. Designers should explicitly consider how spatial perception difficulties affect users' ability to evaluate AI output.

Future Studies. In our study, situational data like queue length was visualized using human-shaped glyphs. A natural possibility is to compare visual encoding (e.g., glyphs vs. numeric labels) to evaluate how data representation like legibility, uncertainty and saliency influences AI reliance.

Additionally, examining how different fundamental characterizations of the situation, such as task difficulty, information ambiguity and time pressure, modulate trust and reliance in AR embedded visualizations, would offer deeper insight into AR-AI interaction design.

6.2 Cognitive Biases Triggered by AR X-Ray Can Induce Inappropriate AI Reliance

Beyond perceptual difficulty, our qualitative findings point to cognitive biases that are unique to the AR X-ray. Participants reported visual proximity illusions and increased trust in AI suggestions due to their spatial embodiment in the environment. These effects align with prior research on spatial cognition biases [71] and trust in embodied agents [43].

Although cognitive biases have been studied in traditional visualization contexts [17, 92], their implications in embodied, spatial, and AI-augmented AR systems remain underexplored. As Padilla *et al.* [64] suggested, when systems are designed with such biases in mind, users can benefit from

Type 1 (intuitive) thinking—particularly under time pressure. Our findings highlight a fertile area for interdisciplinary research at the intersection of spatial cognition, embedded visualizations, and human-AI decision-making.
Design Implications. Embedded visualizations in AR should explicitly account for and possibly counteract spatial cognitive biases. For example, visual proximity illusions may be mitigated through alternative depth cues, or warnings when direct-line views differ from walking paths. Designers might also consider using less embodied, realistic representation for AI suggestions when the risk of over-reliance is high.
Future Studies. While spatial AR studies are often limited by physical constraints of real-world spaces, the increasing accessibility of AR and AI-integrated devices makes spatial cognition experiments increasingly feasible. Future research should investigate how AR-induced biases interact with varying levels of AI accuracy or explanation quality, particularly in high-stakes or time-sensitive scenarios. This represents a promising interdisciplinary frontier spanning visual computing, cognitive science, and human-computer interaction.

6.3 Aligning the Strengths of AR X-Ray with Spatial Tasks

In our study, while the X-ray did not outperform the Minimap on decision accuracy, it significantly reduced post-trial pointing errors, suggesting improved spatial mapping. This distinction suggest that AR X-Ray may be more beneficial in continuous, action-based spatial workflows rather than in isolated decision-making tasks.

In spatial tasks like emergency response, selecting a correct location is only the first step; what follows is action. Misinterpreting a target's location after selection can be consequential. For instance, a user might correctly choose an exit from a wall-mounted map but walk the wrong way due to reference-frame confusion. AR X-Ray mitigates this risk by spatially anchoring information in the environment. Thus, rather than emphasizing the weakness of AR X-Ray, our findings illustrate their potential strengths in embodied spatial tasks and highlight when and where AR embedded visualizations are best applied.
Design Implications. AR embedded visualizations may be most effective when tightly coupled with tasks involving physical action, rather than static decision-making. Designers should consider the viewer's movement and real-time orientation when designing the embedded visualizations. Emerging work on "Visualization in Motion" [104] and "First-Person View Visualization" [37] provides valuable guidance in this direction. In addition, AR systems should support smooth transitions between embedded visualizations and traditional 2D visualizations in the headset depending on task context and user needs.
Future Studies. In our study, participants selected targets verbally. A natural follow-up would involve embodied selection, for example, walking to or pointing at targets, to examine whether the spatial support of AR embedded visualizations becomes more critical under physical interaction. We hypothesize that in such embodied tasks, participants using 2D maps will incur greater cognitive load in spatial mapping, possibly diminishing their ability to evaluate targets effectively. Future work should explore how task modality, such as static vs. embodied, interacts with visualization and AI support in complex environments.

6.4 AR X-Ray May Lower Cognitive Engagement Compared with 2D Visualizations

In our study, there was no significant difference in response time between the `Minimap+NoAI` and `X-ray+NoAI` conditions once initial search time was excluded. However, participants in the `X-ray+AI` condition made decisions more quickly yet achieved lower overall accuracy than those in the `Minimap+AI` condition. This pattern aligns with reliance on heuristic judgments, often described as Type 1 thinking [13, 24]. Together with the high self-reported confidence, these results suggest a lack of deeper cognitive engagement: once AI assistance was available, participants quickly felt confident enough in their intuitive judgments, rather than engaging in more analytical (Type 2) thinking. These findings imply that AR X-Ray, by offering a highly immersive or embodied experience, may inadvertently reduce users' motivation to scrutinize information.
Design Implications. It remains unclear whether this lower cognitive engagement is specific to our task design, a broader characteristic of AR embedded visualizations, or a more general cognitive shift from 2D to AR egocentric perspective. Design strategies that prompt more deliberation, such as "cognitive forcing functions" [10], could be incorporated to encourage deeper analytic thinking in AR environments.
Future Studies. Further research is needed to determine whether AR embedded visualizations consistently reduce cognitive motivation compared to 2D visualizations, and to explore methods for mitigating AR-induced cognitive shortcuts across diverse tasks, visualization designs, and populations.

6.5 Study Limitations

Our study offers an initial examination of AI-aided spatial decision-making with AR embedded visualizations. Although guided by prior research on AR visual systems for spatial tasks [100, 101], we acknowledge that our system does not represent a fully optimized, advanced, or universal solution.

To ensure experimental control, we adopted fixed parameters, such as a discrete set of spatial arrangements, an AI accuracy of 75% (see Section 4.7), and a 20-second decision window, which may not capture the full complexity of high-stakes, real-world environments or more advanced AI systems. Consequently, our findings are most transferable to tasks resembling the indoor, time-sensitive context described in Section 3.1. Moreover, our participant pool primarily comprised well-educated individuals, which may not reflect the behavior of specialized user groups, for example, emergency responders or security professionals. As is common in spatial cognition research [86], our sample size was modest; although sufficient for controlled experimentation, a larger and more diverse sample would strengthen external validity.

Finally, in our analysis, there may be a minor initial search time in the `Minimap` condition, which depends on how quickly participants detect the first two targets in the Minimap. However, this subtle interval cannot be reliably measured from first-person video alone. Future work incorporating eye-tracking data, which is currently unavailable on Apple Vision Pro, could help capture this duration more accurately. We also acknowledge that during S2, some cognitive evaluation of the first target may occur. Nevertheless, by definition, a decision can only be made once the second target has been located. Overall, we view this study as a scientifically grounded first step that provides a reasonable and well-documented foundation for subsequent research to extend and refine AR-based decision support across broader tasks and populations.

7 Conclusion

In this paper, we investigated the impact of AR embedded visualizations on human-AI reliance in spatial decision-making under time constraints by comparing X-ray with Minimap. Our user study (N = 32) revealed that the X-ray visualization tends to evoke over-reliance on AI during spatial decision-making. This effect may be linked to perceptual challenges like occlusion and distance estimation, and AR-specific biases such as visual proximity illusions and the heightened sense of realism and immersion induced by high-fidelity visuals. Nonetheless, the X-ray visualization exhibits unique strengths by enhancing spatial mapping, as evidenced by significantly reduced pointing errors. We hope that our findings and design implications will guide future research and the development of solutions to better leverage AR's advantages to improve human-AI collaboration in spatial contexts.

Acknowledgments

The author(s) would like to express sincere gratitude to the anonymous reviewers for their constructive suggestions.

References

[1] Fadel Adib and Dina Katabi. 2013. See through walls with WiFi! *SIGCOMM Comput. Commun. Rev.* 43, 4 (Aug. 2013), 75–86. doi:10.1145/2534169.2486039

[2] Melanie Bancilhon, Lace Padilla, and Alvitta Ottley. 2023. Evaluating Visualization Decision-Making with Cognitive Models. *Visualization Psychology* (2023).

[3] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance. In *Proc. of CHI (CHI '21)*. ACM, Article 81, 16 pages. doi:10.1145/3411764.3445717

[4] Cindy Xiong Bearfield, Lisanne Van Weelden, Adam Waytz, and Steven Franconeri. 2024. Same data, diverging perspectives: The power of visualizations to elicit competing interpretations. *IEEE TVCG* (2024).

[5] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proc. of CHI.*

[6] Mark Billinghurst, Adrian Clark, and Gun Lee. 2015. *A Survey of Augmented Reality.* now. doi:10.1561/1100000049

[7] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative research in sport, exercise and health* 11, 4 (2019), 589–597.

[8] Nathalie Bressa, Henrik Korsgaard, Aurélien Tabard, Steven Houben, and Jo Vermeulen. 2022. What's the Situation with Situated Visualization? A Survey and Perspectives on Situatedness. *IEEE TVCG* 28, 1 (2022), 107–117. doi:10.1109/IEEETVCG.2021.3114835

[9] Camelia D. Brumar, Sam Molnar, Gabriel Appleby, Kristi Potter, and Remco Chang. 2024. A Typology of Decision-Making Tasks for Visualization. arXiv:2404.08812 [cs.HC] https://arxiv.org/abs/2404.08812

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z. Gajos. 2021. To Trust or to Think: Cognitive Forcing Functions Can Reduce Overreliance on AI in AI-assisted Decision-making. *Proc. ACM Hum.-Comput. Interact.* (2021). doi:10.1145/3449287

[11] Grigore C Burdea and Philippe Coiffet. 2003. *Virtual reality technology.* John Wiley & Sons.

[12] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-based explanations don't help people detect misclassifications of online toxicity. In *Proc. of AAAI.*

[13] Harris Chaiklin. 2012. Thinking Fast and Slow. *Journal of Nervous and Mental Disease* 200 (2012), 826. https://api.semanticscholar.org/CorpusID:168706774

[14] Remco Chang, Mohammad Ghoniem, Robert Kosara, William Ribarsky, Jing Yang, Evan Suma, Caroline Ziemkiewicz, Daniel Kern, and Agus Sudjianto. 2007. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *VAST*. IEEE.

[15] Kurtis Danyluk, Barrett Ens, Bernhard Jenny, and Wesley Willett. 2021. A Design Space Exploration of Worlds in Miniature. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 122, 15 pages. doi:10.1145/3411764.3445098

[16] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *CSN: Business* (2014). https://api.semanticscholar.org/CorpusID:1646733

[17] Evanthia Dimara, Steven Franconeri, Catherine Plaisant, Anastasia Bezerianos, and Pierre Dragicevic. 2018. A task-based taxonomy of cognitive biases for information visualization. *IEEE TVCG* (2018).

[18] Evanthia Dimara and John Stasko. 2022. A Critical Reflection on Visualization Research: Where Do Decision Making Tasks Hide? *IEEE TVCG* (2022). doi:10.1109/IEEETVCG.2021.3114813

[19] Weihua Dong, Yulin Wu, Tong Qin, Xinran Bian, Yan Zhao, Yanrou He, Yawei Xu, and Cheng Yu. 2021. What is the difference between augmented reality and 2D navigation electronic maps in pedestrian wayfinding? *Cartography and Geographic Information Science* (2021).

[20] David Drascic and Paul Milgram. 1996. Perceptual issues in augmented reality. In *Electronic imaging.* https://api.semanticscholar.org/CorpusID:7734379

[21] Mary T. Dzindolet, Linda G. Pierce, Hall P. Beck, and Lloyd A. Dawe. 2002. The Perceived Utility of Human and Automated Aids in a Visual Detection Task. *Human Factors* 44, 1 (2002), 79–94. arXiv:https://doi.org/10.1518/0018720024494856 doi:10.1518/0018720024494856 PMID: 12118875.

[22] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *UIST*. ACM. doi:10.1145/3472749.3474784

[23] Beatrix Emo, Christoph Hoelscher, Jan Malte Wiener, and Ruth Conroy Dalton. 2012. Wayfinding and Spatial Configuration: evidencefrom street corners. In *Wayfinding and Spatial Configuration: evidencefrom street corners.* https://api.semanticscholar.org/CorpusID:128846748

[24] Jonathan St. B. T. Evans and Keith E. Stanovich. 2013. Dual-Process Theories of Higher Cognition: Advancing the Debate. *Perspectives on Psychological Science* (2013). doi:10.1177/1745691612460685

[25] Michael Fernandes, Logan Walls, Sean Munson, Jessica Hullman, and Matthew Kay. 2018. Uncertainty displays using quantile dotplots or cdfs improve transit decision-making. In *Proc. of CHI.*

[26] Aimen Gaba, Zhanna Kaufman, Jason Cheung, Marie Shvakel, Kyle Wm. Hall, Yuriy Brun, and Cindy Xiong Bearfield. 2024. My Model is Unfair, Do People Even Care? Visual Design Affects Trust and Perceived Bias in Machine Learning. *IEEE TVCG* (2024). doi:10.1109/IEEETVCG.2023.3327192

[27] Romeo Giuliano, Franco Mazzenga, Marco Petracca, and Marco Vari. 2013. Indoor localization system for first responders in emergency scenario. *IWCMC* (2013).

[28] Lakshmi Goel, Iris A. Junglas, Blake Ives, and Norman A. Johnson. 2012. Decision-making in-socio and in-situ: Facilitation in virtual worlds. *Decis. Support Syst.* 52 (2012), 342–352. https://api.semanticscholar.org/CorpusID:46128169

[29] Jeffrey R. Guerrieri, Michael H. Francis, P. F. Wilson, T. Kos, L. E. Miller, Nelson P. Bryner, D. W. Stroup, and L T. Klein-berndt. 2006. RFID-assisted indoor localization and communication for first responders. *2006 First European Conference on Antennas and Propagation* (2006), 1–6. https://api.semanticscholar.org/CorpusID:16040353

[30] Ziyang Guo, Yifan Wu, Jason D Hartline, and Jessica Hullman. 2024. A decision theoretic framework for measuring AI reliance. In *Proc. of FAccAT.*

[31] Sunwoo Ha, Shayan Monadjemi, and Alvitta Ottley. 2024. Guided By AI: Navigating Trust, Bias, and Data Exploration in AI-Guided Visual Analytics. In *Computer Graphics Forum.* Wiley Online Library.

[32] Justin Harris, Kathy Hirsh-Pasek, and Nora S. Newcombe. 2013. Understanding spatial transformations: similarities and differences between mental rotation and mental folding. *Cognitive Processing* (2013). https://api.semanticscholar.org/CorpusID:6072708

[33] Mordechai I Henig and John T Buchanan. 1996. Solving MCDM problems: Process concepts. *Journal of multi-criteria decision analysis* (1996).

[34] Teresa Hirzle, Florian Müller, Fiona Draxler, Martin Schmitz, Pascal Knierim, and Kasper Hornbæk. 2023. When XR and AI Meet - A Scoping Review on Extended Reality and Artificial Intelligence. In *Proc. of CHI.* ACM. doi:10.1145/3544548.3581072

[35] Kevin Anthony Hoff and Masooda Bashir. 2015. Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust. *Human Factors* 57, 3 (2015), 407–434. arXiv:https://doi.org/10.1177/0018720814547570 doi:10.1177/0018720814547570 PMID: 25875432.

[36] Sathaporn Hu, Joseph Malloch, and Derek Reilly. 2021. A Comparative Evaluation of Techniques for Locating Out of View Targets in Virtual Reality. In *Graphics Interface 2021.*

[37] Petra Isenberg, Lionel Reveret, and Romain Vuillemot. 2024. HomeOlympics: Turning Everyday Outdoor Physical Activities into Olympic-Style Experience. In *Proc. of the IEEE VIS Workshop on First-Person Visualizations for Outdoor Physical Activities.* Virtual, Unknown Region. https://inria.hal.science/hal-04782928

[38] Mikkel R. Jakobsen and Kasper Hornbæk. 2013. Interactive Visualizations on Large and Small Displays: The Interrelation of Display Size, Information Space, and Scale. *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2336–2345. doi:10.1109/TVCG.2013.170

[39] Yvonne Jansen, Pierre Dragicevic, Petra Isenberg, Jason Alexander, Abhijit Karnik, Johan Kildal, Sriram Subramanian, and Kasper Hornbæk. 2015. Opportunities and Challenges for Data Physicalization. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/2702123.2702180

[40] Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. 2018. To trust or not to trust a classifier. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18)*. Curran Associates Inc., 5546–5557.

[41] Alex Kale, Matthew Kay, and Jessica Hullman. 2020. Visual reasoning strategies for effect size judgments and decisions. *IEEE TVCG* (2020).

[42] Denis Kalkofen, Erick Mendez, and Dieter Schmalstieg. 2009. Comprehensible Visualization for Augmented Reality. *IEEE Transactions on Visualization and Computer Graphics* 15, 2 (2009), 193–204. doi:10.1109/TVCG.2008.96

[43] Kangsoo Kim, Luke Boelling, Steffen Haesler, Jeremy Bailenson, Gerd Bruder, and Greg F. Welch. 2018. Does a Digital Assistant Need a Body? The Influence of Visual Embodiment and Social Behavior on the Perception of Intelligent Virtual Agents in AR. In *Proc. ISMAR.* doi:10.1109/ISMAR.2018.00039

[44] Roberta L. Klatzky. 1998. Allocentric and Egocentric Spatial Representations: Definitions, Distinctions, and Interconnections. In *Spatial Cognition.* https://api.semanticscholar.org/CorpusID:7264069

[45] Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*

15, 2 (2016), 155–163.

[46] Julian Kreiser, Alexander Hann, Eugen Zizer, and Timo Ropinski. 2017. Decision graph embedding for high-resolution manometry diagnosis. *IEEE TVCG* (2017).

[47] J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics* (1977), 159–174.

[48] David Launder and Chad Perry. 2014. A study identifying factors influencing decision making in dynamic emergencies like urban fire and rescue settings. In *A study identifying factors influencing decision making in dynamic emergencies like urban fire and rescue settings.* https://api.semanticscholar.org/CorpusID:110717122

[49] Connor Lawless, Jakob Schoeffer, Lindy Le, Kael Rowan, Shilad Sen, Cristina St. Hill, Jina Suh, and Bahareh Sarrafzadeh. 2024. "I Want It That Way": Enabling Interactive Decision Support Using Large Language Models and Constraint Programming. *TIIS* (2024). doi:10.1145/3685053

[50] Benjamin Lee, Michael Sedlmair, and Dieter Schmalstieg. 2023. Design patterns for situated visualization in augmented reality. *IEEE TVCG* (2023).

[51] Chunggi Lee, Tica Lin, Hanspeter Pfister, and Chen Zhu-Tian. 2024. Sportify: question answering with embedded visualizations and personified narratives for sports video. *IEEE Transactions on Visualization and Computer Graphics* (2024).

[52] Jaewook Lee, Fanjie Jin, Younsoo Kim, and David Lindlbauer. 2022. User Preference for Navigation Instructions in Mixed Reality. In *IEEE VR*. 802–811. doi:10.1109/VR51125.2022.00102

[53] Jiahao Nick Li, Yan Xu, Tovi Grossman, Stephanie Santosa, and Michelle Li. 2024. OmniActions: Predicting Digital Actions in Response to Real-World Multimodal Sensory Inputs with LLMs. In *Proc. of CHI*. ACM. doi:10.1145/3613904.3642068

[54] Zhenlong Li and Huan Ning. 2023. Autonomous GIS: the next-generation AI-powered GIS. arXiv:2305.06453 [cs.AI] https://arxiv.org/abs/2305.06453

[55] Tica Lin, Alexandre Aouididi, Chen Zhu-Tian, Johanna Beyer, Hanspeter Pfister, and Jui-Hsien Wang. 2023. VIRD: Immersive Match Video Analysis for High-Performance Badminton Coaching. *IEEE TVCG* (2023). https://api.semanticscholar.org/CorpusID:260125958

[56] Xiaoan Liu, Difan Jia, Xianhao Carton Liu, Mar Gonzalez-Franco, and Chen Zhu-Tian. 2025. Reality Proxy: Fluid Interactions with Real-World Objects in MR via Abstract Representations. *arXiv preprint arXiv:2507.17248* (2025).

[57] Mark A. Livingston, Arindam Dey, Christian Sandor, and B. Thomas. 2013. Pursuit of "X-Ray Vision" for Augmented Reality. In *Pursuit of "X-Ray Vision" for Augmented Reality.* https://api.semanticscholar.org/CorpusID:16721214

[58] Helena Löfström. 2023. On the Definition of Appropriate Trust and the Tools that Come with it. *2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE)* (2023), 1555–1562. https://api.semanticscholar.org/CorpusID:262084200

[59] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in ai-assisted decision-making. In *Proc. of CHI*.

[60] G. Manfredi, Nicola Felice Capece, R. P. Di, and Ugo Erra. 2024. A Mixed Reality Application for Multi-Floor Building Evacuation Drills using Real-Time Pathfinding and Dynamic 3D Modeling. https://api.semanticscholar.org/CorpusID:273970950

[61] Katelyn Morrison, Donghoon Shin, Kenneth Holstein, and Adam Perer. 2023. Evaluating the impact of human explanation strategies on human-AI visual decision-making. *Proc. of HCI* (2023).

[62] R Nydegger, L Nydegger, and F Basile. 2011. Post-traumatic stress disorder and coping among career professional firefighters. *American Journal of Health Sciences* (2011).

[63] Emre Oral, Ria Chawla, Michel Wijkstra, Narges Mahyar, and Evanthia Dimara. 2023. From Information to Choice: A Critical Inquiry Into Visualization Tools for Decision Making. *IEEE TVCG* (2023). doi:10.1109/IEEETVCG.2023.3326593

[64] Lace M Padilla, Sarah H Creem-Regehr, Mary Hegarty, and Jeanine K Stefanucci. 2018. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive research: principles and implications* (2018).

[65] Lace M. K. Padilla, Sarah H. Creem-Regehr, Mary Hegarty, and Jeanine K. Stefanucci. 2018. Decision making with visualizations: a cognitive framework across disciplines. *Cognitive Research: Principles and Implications* (2018).

[66] Stephan Pajer, Marc Streit, Thomas Torsney-Weir, Florian Spechtenhauser, Torsten Möller, and Harald Piringer. 2016. Weightlifter: Visual weight space exploration for multi-criteria decision making. *IEEE TVCG* (2016).

[67] Greg Penney, David Launder, Joe Cuthbertson, and Matthew B Thompson. 2022. Threat assessment, sense making, and critical decision-making in police, military, ambulance, and fire services. *Cognition, Technology & Work* (2022).

[68] Leon Pietschmann, Paul-David Joshua Zuercher, Erik Bub'ik, Chen Zhu-Tian, Hanspeter Pfister, and Thomas Bohné. 2023. Quantifying the Impact of XR Visual Guidance on User Performance Using a Large-Scale Virtual Assembly Experiment. *IEEE VIS* (2023). https://api.semanticscholar.org/CorpusID:260680537

[69] George Pu, Paul Wei, Amanda Aribe, James Boultinghouse, Nhi Dinh, Fang Xu, and Jing Du. 2021. Seeing Through Walls: Real-Time Digital Twin Modeling of Indoor Spaces. In *2021 Winter Simulation Conference (WSC)*. doi:10.1109/WSC52266.2021.9715364

[70] Carlos Quijano-Chavez, Nina Doerr, Benjamin Lee, Dieter Schmalstieg, and Michael Sedlmair. 2024. Brushing and Linking for Situated Analytics. In *2024 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, 597–603.

[71] Priya Raghubir and Aradhna Krishna. 1996. As the crow flies: Bias in consumers' map-based distance judgments. *Journal of Consumer Research* (1996).

[72] Nils Reisen, Ulrich Hoffrage, and Fred W Mast. 2008. Identifying decision strategies in a consumer choice situation. *Judgment and decision making* (2008).

[73] Jonatan Reyes, Anil Ufuk Batmaz, and Marta Kersten-Oertel. 2025. Trusting AI: does uncertainty visualization affect decision-making? *Frontiers in Computer Science* (2025).

[74] J Edward Russo and France Leclerc. 1994. An eye-fixation analysis of choice processes for consumer nondurables. *Journal of consumer research* (1994).

[75] David Saffo, Sara Di Bartolomeo, Tarik Crnovrsanin, Laura South, Justin Raynor, Caglar Yildirim, and Cody Dunne. 2024. Unraveling the Design Space of Immersive Analytics: A Systematic Review. *IEEE TVCG* 30, 1 (2024), 495–506. doi:10.1109/IEEETVCG.2023.3327368

[76] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with Uncertainty: Understanding the Impact of Prognostic Versus Diagnostic Tasks on Trust and Reliance in Human-AI Decision Making. In *Proc. of CHI*. ACM. doi:10.1145/3613904.3641905

[77] Max Schemmer, Niklas Kuehl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. In *Proc. of IUI*. ACM. doi:10.1145/3581641.3584066

[78] Sharad Sharma, James Stigall, and Sri Teja Bodempudi. 2020. Situational Awareness-based Augmented Reality Instructional (ARI) Module for Building Evacuation. In *VRW*. doi:10.1109/VRW50115.2020.00020

[79] Sungbok Shin, Andrea Batch, Peter William Scott Butcher, Panagiotis D. Ritsos, and Niklas Elmqvist. 2024. The Reality of the Situation: A Survey of Situated Analytics. *IEEE TVCG* (Aug 2024). doi:10.1109/IEEETVCG.2023.3285546

[80] Alexander W. Siegel and Sheldon H. White. 1975. The development of spatial representations of large-scale environments. *Advances in child development and behavior* (1975). https://api.semanticscholar.org/CorpusID:28635993

[81] Herbert A. Simon. 1977. *The Logic of Heuristic Decision Making*. Springer Netherlands. doi:10.1007/978-94-010-9521-1_10

[82] Gregorio Soria, Lidia M. Ortega-Alvarado, and Francisco Ramón Feito-Higueruela. 2018. Augmented and Virtual Reality for Underground Facilities Management. *J. Comput. Inf. Sci. Eng.* (2018). https://api.semanticscholar.org/CorpusID:116308226

[83] Richard Stoakley, Matthew J. Conway, and Randy Pausch. 1995. Virtual reality on a WIM: interactive worlds in miniature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '95)*. ACM Press/Addison-Wesley Publishing Co., USA, 265–272. doi:10.1145/223904.223938

[84] Xia Su, Han Zhang, Kaiming Cheng, Jaewook Lee, Qiaochu Liu, Wyatt Olson, and Jon E. Froehlich. 2024. RASSAR: Room Accessibility and Safety Scanning in Augmented Reality. In *Proc. of CHI*. ACM. doi:10.1145/3613904.3642140

[85] Siddharth Swaroop, Zana Buçinca, Krzysztof Z. Gajos, and Finale Doshi-Velez. 2024. Accuracy-Time Tradeoffs in AI-Assisted Decision Making under Time Pressure. In *IUI*. ACM. doi:10.1145/3640543.3645206

[86] Haoyu Tan, Tongyu Nie, and Evan Suma Rosenberg. 2024. Invisible Mesh: Effects of X-Ray Vision Metaphors on Depth Perception in Optical-See-Through Augmented Reality. *IEEE VR* (2024). https://api.semanticscholar.org/CorpusID:269175599

[87] Barbara Tversky. 2003. Structures Of Mental Spaces: How People Think About Space. *Environment and Behavior* (2003). doi:10.1177/0013916502238865

[88] Barbara Tversky. 2008. *Spatial Cognition: Embodied and Situated*. Cambridge University Press.

[89] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S. Bernstein, and Ranjay Krishna. 2023. Explanations Can Reduce Overreliance on AI Systems During Decision-Making. *Proc. ACM Hum.-Comput. Interact.* CSCW1 (2023). doi:10.1145/3579605

[90] Emily Wall, Subhajit Das, Ravish Chawla, Bharath Kalidindi, Eli T Brown, and Alex Endert. 2017. Podium: Ranking data using mixed-initiative visual analytics. *IEEE TVCG* (2017).

[91] Emily Wall, Laura Matzen, Mennatallah El-Assady, Peta Masters, Helia Hosseinpour, Alex Endert, Rita Borgo, Polo Chau, Adam Perer, Harald Schupp, et al. 2024. Trust Junk and Evil Knobs: Calibrating Trust in AI Visualization. In *PacificVis*. IEEE.

[92] Emily Wall, Arpit Narechania, Adam Coscia, Jamal Paden, and Alex Endert. 2021. Left, right, and gender: Exploring interaction traces to mitigate human biases. *IEEE TVCG* (2021).

[93] Qianwen Wang, Kexin Huang, Payal Chandak, Marinka Zitnik, and Nils Gehlenborg. 2022. Extending the nested model for user-centric xai: A design study on gnn-based drug repurposing. *IEEE TVCG* (2022).

[94] Amelia C Warden, Christopher D Wickens, Domenick Mifsud, Shannon Ourada, Benjamin A Clegg, and Francisco R Ortega. 2022. Visual search in augmented reality: Effect of target cue type and location. In *Proceedings of the human factors*

and ergonomics society annual meeting, Vol. 66. SAGE Publications Sage CA: Los Angeles, CA, 373–377.

[95] Di Weng, Heming Zhu, Jie Bao, Yu Zheng, and Yingcai Wu. 2018. Homefinder revisited: Finding ideal homes with reachability-centric multi-criteria decision making. In *Proc. of CHI*.

[96] Jan Malte Wiener, Simon J. Büchner, and Christoph Hölscher. 2009. Taxonomy of Human Wayfinding Tasks: A Knowledge-Based Approach. *Spatial Cognition & Computation* (2009). https://api.semanticscholar.org/CorpusID:16529538

[97] Wesley Willett, Yvonne Jansen, and Pierre Dragicevic. 2017. Embedded Data Representations. *IEEE TVCG* (2017). doi:10.1109/IEEETVCG.2016.2598608

[98] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. of CHI*. ACM. doi:10.1145/1978942.1978963

[99] Tim Wächter, Jan Rexilius, and Matthias König. 2022. Interactive evacuation in intelligent buildings assisted by mixed reality. *Journal of Smart Cities and Society* (2022). arXiv:https://journals.sagepub.com/doi/pdf/10.3233/SCS-220009 doi:10.3233/SCS-220009

[100] Fang Xu, Tianyu Zhou, Tri Nguyen, and Jing Du. 2024. Augmented reality in team-based search and rescue: Exploring spatial perspectives for enhanced navigation and collaboration. *Safety Science* (2024). doi:10.1016/j.ssci.2024.106556

[101] Fang Xu, Tianyu Zhou, Hengxu You, and Jing Du. 2024. Improving indoor wayfinding with AR-enabled egocentric cues: A comparative study. *Advanced Engineering Informatics* 59 (2024), 102265. doi:10.1016/j.aei.2023.102265

[102] Liuchang Xu, Shuo Zhao, Qingming Lin, Luyao Chen, Qianqian Luo, Sensen Wu, Xinyue Ye, Hailin Feng, and Zhenhong Du. 2024. Evaluating Large Language Models on Spatial Tasks: A Multi-Task Benchmarking Study. *ArXiv* (2024). https://api.semanticscholar.org/CorpusID:271956990

[103] Fumeng Yang, Zhuanyi Huang, Jean Scholtz, and Dustin L Arendt. 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In *Proc. of IUI*.

[104] Lijie Yao, Anastasia Bezerianos, Romain Vuillemot, and Petra Isenberg. 2022. Visualization in Motion: A Research Agenda and Two Evaluations. *IEEE Trans. Vis. Comput. Graph.* (2022). doi:10.1109/IEEETVCG.2022.3184993

[105] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proc. of CHI*. ACM. doi:10.1145/3290605.3300509

[106] Kexin Zhang, Brianna R Cochran, Ruijia Chen, Lance Hartung, Bryce Sprecher, Ross Tredinnick, Kevin Ponto, Suman Banerjee, and Yuhang Zhao. 2024. Exploring the Design Space of Optical See-through AR Head-Mounted Displays to Support First Responders in the Field. In *Proc. of CHI (CHI '24)*. ACM. doi:10.1145/3613904.3642195

[107] Jieqiong Zhao, Yixuan Wang, Michelle V Mancenido, Erin K Chiou, and Ross Maciejewski. 2023. Evaluating the impact of uncertainty visualization on model reliance. *IEEE TVCG* (2023).

[108] Chen Zhu-Tian, Daniele Chiappalupi, Tica Lin, Yalong Yang, Johanna Beyer, and Hanspeter Pfister. 2024. RL-LABEL: A Deep Reinforcement Learning Approach Intended for AR Label Placement in Dynamic Scenarios. *TVCG* (2024). doi:10.1109/IEEETVCG.2023.3326568

[109] Chen Zhu-Tian, Qisen Yang, Jiarui Shan, Tica Lin, Johanna Beyer, Haijun Xia, and Hanspeter Pfister. 2023. iBall: Augmenting Basketball Videos with Gaze-Moderated Embedded Visualizations. In *Proc. of CHI*.

[110] Chen Zhu-Tian, Qisen Yang, Xiao Xie, Johanna Beyer, Haijun Xia, Yingcai Wu, and Hanspeter Pfister. 2023. Sporthesia: Augmenting Sports Videos Using Natural Language. *IEEE Trans. Vis. Comput. Graph.* 29, 1 (2023), 918–928. doi:10.1109/TVCG.2022.3209497