Working Easier or Working Harder? The Impact of Cognitive Load on Synergy in Human–AI Collaboration

Jiaxin Zhang

MSc Thesis
Leiden University
Faculty of Science
Creative Intelligence and Technology

Supervisor: Dr. Max van Duijn
Second Supervisor: Dr. Michiel van der Meer

January 16, 2026

Contents

1 Introduction

Imagine the utopia of Human-AI collaboration: Humans leverage creativity, intuition, and contextual understanding, while AI brings speed, scalability, and analytical power, enabling better decision-making together (Vaccaro et al., 2024). Such forms of collaborative intelligence are already being adopted in fields such as medicine(Topol, 2019), finance (Brynjolfsson et al., 2017), manufacturing (Lee et al., 2018), and education (Holmes et al., 2019). Human-AI collaboration is founded on the principle that AI and human expertise are different – one excels in areas where the other may be limited. This minimal overlap in errors creates opportunities for one to compensate for the other's mistakes, a concept referred to as complementarity in hybrid intelligence (Akata et al., 2020) and human-AI collaboration (Bansal et al., 2021).

However, extensive research has shown that despite the fact that the AI system did on average help humans perform better, Human–AI teams often do not outperform AI-only teams, suggesting a low synergy in Human-AI team (Vaccaro et al., 2024). These suboptimal outcomes can be attributed to a range of factors, broadly categorized into four groups: collaboration characteristics, task characteristics, AI characteristics, and human characteristics (Hemmer et al., 2021).

While previous efforts in improving Human-AI collaboration have primarily centered on enhancing AI systems - such as improving model accuracy and enhancing explainability - these interventions have shown limited success in significantly boosting complementarity team performance(CTP). For instance, simply presenting LLM-based analyses to users either randomly or concurrently does not increase their AI-assisted decision accuracy (Li et al., 2025). Although explanation can increase user trust, they fail to lead to better decisions (Bansal et al., 2021). In contrast, modeling how humans interpret and respond to AI-generated analyses, rather than treating explanations as static information, can foster greater human–AI complementarity (Li et al., 2025). Such responses are shaped by users' underlying cognitive processes, achieving genuine complementarity requires understanding how individuals differ in cognitive ability and reasoning profiles when collaborating with AI (Hemmer et al., 2021).

Through this study, I aim to understand how **human factors**, particularly **cognitive load** and **behavioral choices**, shape the dynamics of human-AI collaboration. Specifically, when individuals experience varying levels of cognitive load, how do they adapt their preferred collaboration strategies? Do they prefer the AI to adapt to them and follow their lead, or do they prefer to AI to take the lead and instruct them what to do? Moreover, do certain strategy choices lead to better outcomes under specific cognitive conditions? Accordingly, this study addresses the following research questions:

RQ1: How does cognitive load influence users' preference for collaboration strategy (AI-led vs. human-led)?

RQ2: How does cognitive load influence performance in Human–AI teams?

RQ3: What conditions lead to better individual performance in Human–AI collaboration?

To better simulate the time pressure and dynamic nature of real-world collaborative scenarios, we adopt the Overcooked framework (Zhang et al., 2025). Unlike turn-by-turn collaborative tasks, simultaneous collaboration tasks that are time sensitive require real-time responses to partners and interaction with the environment, as well as reasoning about dynamically changing human partners' strategies and environments (Zhang et al., 2025).

To capture individual cognitive states and behavioral strategies, the study adopts a mixed-methods approach to collect data. Cognitive states are assessed using objective cognitive load measurements, including gaze-based physiological indicators(e.g., pupil dilation) tracked with

the GazePoint3 eye tracker, as well as subjective cognitive load measurements obtained through questionnaires. These combined measures will enable a more detailed analysis of how individual differences shape collaborative efficiency and complementarity in real-time human–AI interaction.

This study offers insights into designing cognitively adaptive collaborative systems (e.g., systems that switch modes based on user state); Furthermore, it helps identify human preferences and performance limits under constrained cognitive resources, guiding adaptive agent behavior.

2 Literature Review

2.1 The evolving landscape of Human-AI collaboration

Different levels of human–AI collaboration depend fundamentally on the levels of automation (Parasuraman et al., 2000). At different automation levels, systems vary in how they support information acquisition, information analysis, decision and action selection, and action implementation. As the capabilities and autonomy of AI systems increase, the nature of the human–AI relationship evolves across qualitatively different roles.

In the earliest stage, AI systems functioned primarily as auxiliary tools, performing fixed, rule-based tasks characteristic of the mechanical and early information-processing eras (Fitts, 1951). With advances in machine learning and predictive modeling, AI systems progressed to the role of adaptive assistants, capable of generating recommendations, anticipating user needs, and autonomously executing parts of a task. Beyond assistance, these systems increasingly participate in shared goal pursuit, decision-making, and task coordination, exhibiting core characteristics of teamwork. The transition from machines as tools, machine as assistants to machines as teammates represents a fundamental paradigm shift in human–machine interaction (Seeber et al., 2020). These systems not only perform tasks independently but can also initiate actions, negotiate objectives, and adapt to dynamic environments.

One intuitive path to improving collaboration/achieve a reliable teammate is simply to build stronger algorithms or more capable agents. The evolution of AI system design from rule-based automation to LLM-based agent architectures has enabled more flexible implementations of reasoning, memory, reflection, autonomous learning, and generalization(Huang et al., 2024). However, prior work suggests that increasing model capability alone does not resolve the core challenges of human–AI teaming. Carroll et al. (2019b) showed that agents trained through self-play or population-based reinforcement learning can coordinate well with other agents but perform poorly when paired with humans. This limitation persists even in modern LLM-based systems: recent studies report that larger or more capable LLMs do not automatically produce better coordination with human partners unless the model explicitly learns or infers human mental model, including mental state, beliefs, intentions and goals, as well as social factors such as trust, adaptability, and error sensitivity(Wiltgen et al., 2024; Fügener et al., 2022).

2.2 Cognitive load in Human-AI Collaboration

Cognitive load theory, introduced by Sweller discussed the cognitive load as a mental resources a person has available for solving problems or completing tasks. Given the very limited short memory capacity of memory, any task that require a large amount of information storage in working memory can lead to high cognitive load (Sweller, 1988). In order to achieve better performance, need to find a balance between/decrease intrinsic and extraneous cognitive load

(Sweller et al., 1998) It was originally developed within educational psychology. As systems became more interactive and less mechanical, understanding mental workload became essential for predicting usability, user errors, and task performance.

Paradoxically, automated systems can both reduce and increase mental workload. Extremes of mental workload can create states of overload or underload, both of which may be detrimental to performance (Wilson and Rajan, 1995). The idea of an optimal level of workload is grounded in attentional resource theory, which proposes that both overload and underload can induce psychological strain due to a mismatch between task demands and human capabilities (Byrne and Parasuraman, 1996; Gopher and Kimchi, 1989). It is increasingly accepted that optimal performance occurs when task demands are appropriately calibrated (Hancock and Caird, 1993).

Conversely, individuals who are prone to stress or fatigue may experience poorer performance under conditions of underload, as they fail to mobilize sufficient compensatory effort to meet task demands (Desmond and Hancock, 1998). Underload has also been associated with passivity, suggesting that optimal workload reflects a need to maintain a certain level of engagement and control (Hockey et al., 1989).

Although cognitive load has long been a central construct in HCI for explaining user performance, attention, and error (e.g., Young and Stanton, 2002; Wickens, 2008), research on human–AI collaboration has largely focused on improving algorithmic performance rather than understanding human cognitive states. Existing human-factors work in human–agent teaming has primarily examined trust (Hancock et al., 2011), sense of agency (Berberian et al., 2012), and workload in classical automation contexts (Parasuraman and Riley, 1997). Yet, the role of cognitive load in human–AI teaming remains largely unexplored, especially in interactive, collaborative settings where humans and AI jointly plan and adapt. This study addresses this gap by examining how cognitive load shapes human–AI collaboration and team performance.

2.3 Measurements in Cognitive load

Cognitive load in Human–AI collaboration is commonly assessed using a combination of subjective and physiological measures, each capturing different facets of mental effort. Prior research consistently shows that subjective ratings and physiological indicators are correlated but not interchangeable, because they operate on different temporal and psychological levels (Rubio et al., 2004). Subjective measures, such as NASA-TLX, provide post-task, global judgments of perceived workload and are easy to administer, yet they depend on memory, interpretation, and self-awareness. In contrast, physiological measures provide continuous moment-to-moment indices of cognitive effort that can capture transient fluctuations invisible to retrospective reports (Yuksel et al., 2016).

Among physiological methods, eye-tracking is considered one of the most promising non-invasive indicators of cognitive load, offering high temporal resolution and minimal disruption to task performance. A central eye-tracked metric is pupil diameter, whose systematic dilation in response to increased mental effort is captured by the well-established Task-Evoked Pupillary Response (TEPR) (Beatty, 1982). Extensive empirical work demonstrates a positive correlation between pupil size and mental workload across diverse environments and task types, including gaming (Sevcenko et al., 2021), virtual environments (van der Wel and van Steenbergen, 2018), driving (Recarte and Nunes, 2000), and controlled cognitive processing tasks (Beatty, 1982).

However, pupil measurements must be interpreted with attention to individual differences (e.g. baseline pupil size, lighting sensitivity) and contextual factors such as luminance, fatigue, and emotional arousal(Laeng et al., 2012). To enhance reliability, studies typically compute baseline-corrected pupil diameter, measuring changes relative to an individual's mean pupil size

recorded during a neutral baseline period. This approach, recommended by Krejtz et al. (2018), isolates cognitive-induced dilation from natural variability and environmental confounding. In this study, cognitive load is therefore assessed using a multi-modal measurement strategy, combining subjective workload ratings with continuous pupil-based physiological indices that also relies on individual baselines, allowing for a more comprehensive and temporally precise understanding of how cognitive load evolves during Human–AI collaboration and how it shapes user' decision-making, agency and task performance.

2.4 Game in HAC

As Human–AI collaboration moves toward increasingly complex and dynamic scenarios, AI systems are no longer conceptualized merely as decision-support tools but as autonomous agents capable of planning, executing, and revising actions at the task level (Seeber et al., 2020). However, many existing experimental paradigms rely on simplified interaction settings, such as one-shot recommendations or static information provision, that fail to capture the coordination, shared situation awareness, mutual prediction in joint task settings (Johnson et al., 2014). In contrast, game environments offer a unique research context in which AI agents can function as independent collaborators, continuously interacting with human teammates under shared goals and constraints.

Recent advances in large language models have further amplified the suitability of game environments for studying Human–AI collaboration, as these models enable agents to operate in ways that more closely resemble real-world collaborative settings. For example, Voyager demonstrates how an LLM-driven embodied agent can autonomously explore, learn, and iteratively refine complex skills in an open-ended Minecraft environment through planning, execution, and self-reflection (Wang et al., 2023). Similarly, Park et al. (2023) introduce Generative Agents, in which LLM-powered agents inhabit a simulated sandbox world, maintaining long-term memory, forming plans, and engaging in socially coherent interactions with other agents, thereby exhibiting emergent behaviors such as coordination, communication, and collective activity.

Among existing game-based benchmarks, Overcooked has been widely validated as an effective framework for studying Human–AI collaboration. In this cooperative cooking game, human and AI teammates must coordinate under time constraints to prepare ingredients, assemble dishes, serve orders, and manage dynamic hazards such as overcooking and spatial bottlenecks. Prior work has established Overcooked as a benchmark for evaluating collaborative AI, demonstrating that agents trained through self-play can achieve high task performance while still failing to coordinate effectively with human partners (Carroll et al., 2019a). Subsequent research has leveraged Overcooked to study human-aware planning (Ho et al., 2019), Theory-of-Mind-based agent modeling(Wu et al., 2021), and human–AI co-adaptation over repeated interactions (Hu et al., 2022). Importantly, the availability of an open-source Overcooked-AI framework enables reproducible experimentation and systematic investigation of human factors in human–AI teamwork.

3 Task Design

To investigate Human–AI collaboration under controlled conditions, this study employs an Overcooked-based collaborative game environment adapted from the framework proposed by Zhang et al. (2025). In the original framework, a shared workspace environment derived from

Overcooked-AI was used to evaluate real-time simultaneous human–AI interaction, including macro actions, task rewards, and state access for both agent and human players collaborating to complete cooking tasks under time pressure.

Building on this foundation, we extended the environment in three key respects. First, whereas the original implementation focused on single layouts for simultaneous interaction, we introduced two distinct collaborative rounds that explicitly differentiate AI-led and human-led control modes: In AI-led mode, the AI provides task directives, whereas in human-led mode, the human takes the lead and the AI follows; Second, we integrated a real-time mode-switching interface allowing participants to toggle between collaboration modes during gameplay; Third, the original agent was replaced with an LLM-driven adaptive AI agent whose behavior adapts dynamically according to the selected control mode.

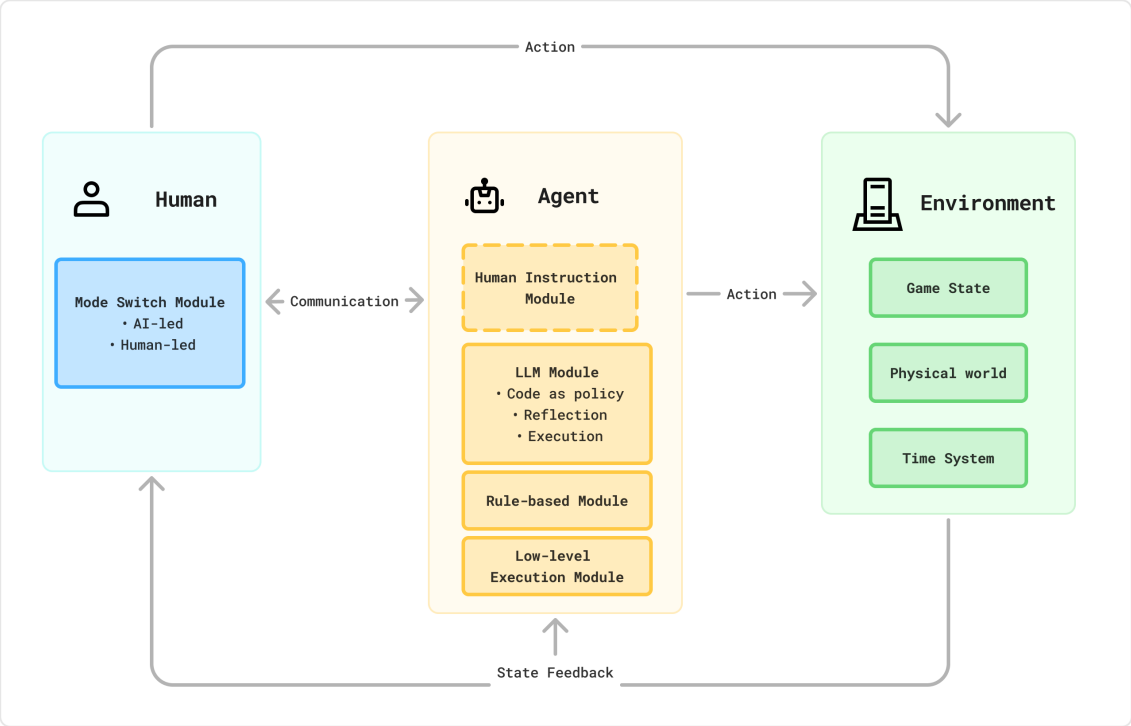

Figure 1: The diagram shows the interaction loop between the human, the AI agent, and the environment in an Overcooked-style task. The human selects between AI-led and human-led modes, while the AI agent processes instructions and executes actions through LLM-based and rule-based modules. Both interact with a shared environment that provides continuous state feedback for adaptive collaboration.

3.1 Human

The task incorporates a mode-switching module that supports two distinct collaboration modes:

- **AI-led mode:** The agent actively analyzes the game state, generates task allocations, and communicates these assignments to the human, who then cooperates in their execution.
- **Human-led mode:** The human issues instructions, which the agent interprets and subsequently assists in executing.

3.2 Agent

The agent is designed with four layers, realizing the complete intelligence flow from high-level strategy (macro-actions) to corresponding action execution (micro-actions). This layered structure ensures both real-time responsiveness and intelligent decision-making capabilities.

- **Human Instruction Module:** In the human-led mode, this module interprets and executes tasks assigned by the human. It incorporates an order-mapping table, which enables the translation of textual instructions (e.g., "prepare beef") into corresponding macro-actions (e.g., (*prepare*, {food: "Beef"})).
- **LLM Module:** Serving as the core intelligence of the system, this module employs an asynchronous invocation mechanism, triggered every 75 time steps. It performs reflective strategy updates based on historical records. Following prior research indicating strong performance of GPT models, we adopt `gpt-3.5-turbo`, which ensures both responsiveness and quality. The module provides two key functionalities:
  1. *Code-as-Policy:* Generates conditional macro-actions in the form of (*condition*, *action*) pairs, as well as order priorities represented as ranked lists of strings.
  2. *Reflection:* Produces reflective inputs based on historical trajectories, updating behavioral guidelines and inferring human behavior patterns, thereby optimizing task allocation.
- **Rule-based Module:** This module provides a stable foundation for decision-making and includes the following core components:
  - *Action Patterns* (`action_patterns`): Stores predefined action patterns.
  - *Order Processing Patterns* (`order_patterns`): Defines standardized workflows for order handling.
  - *Food-to-Ingredient Mapping* (`food_to_ingredients`): Maps food items to the required ingredients.
  - *Valid Order Verification* (`valid_orders`): Ensures the validity of incoming orders.
  - *Task Update Function* (`update_assignments()`): Updates and maintains the list of assigned tasks.
- **Low-level Execution Module:** This module bridges high-level strategies with executable actions, ensuring real-time interaction with the environment. Its key functions include:
  - *Action Conversion:* Translates high-level instructions into atomic actions.
  - *Path Planning:* Implements path-planning algorithms for efficient navigation.
  - *State Management:* Maintains and updates the agent's internal and environmental states.
  - *Physical Interaction:* Handles interactions with the environment at the level of physical execution.

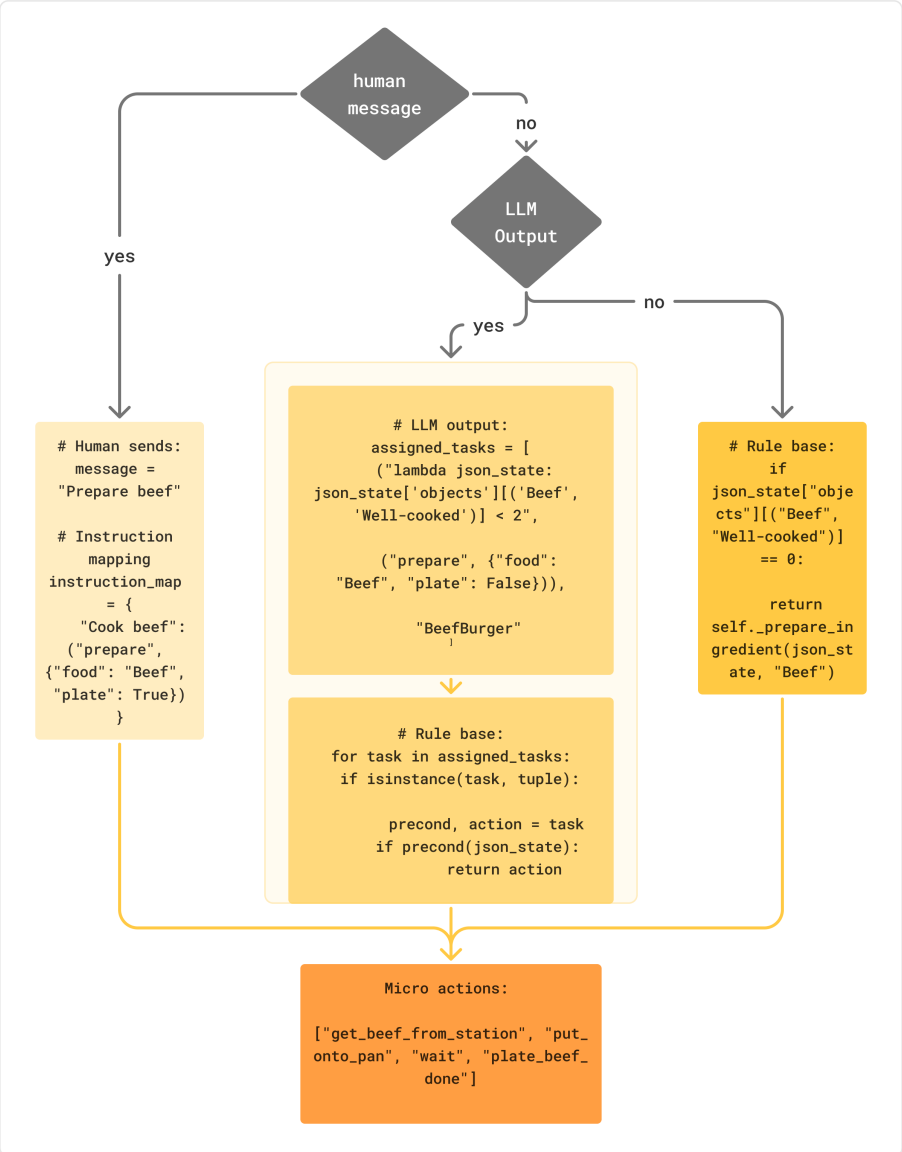

Figure 2: Example of agent orchestration and reaction mechanism.

3.3 Environment

The environment in which the agent operates is composed of three primary subsystems: the *game state*, the *physical world*, and the *time system*. Each subsystem provides essential information for both human and agent decision-making.

- **Game State:** Represents the current status of the game entities and overall progress. It consists of:
  - *Objects:* The state and quantity of game entities.
  - *Orders:* The order management system.
  - *Score:* The scoring system reflecting team performance.
- **Physical World:** Models the interactive environment where cooking and delivery take place. It includes:
  - *Cooking Stations:* Functional stations for food preparation and cooking.
  - *Counters:* Countertop areas used for temporary placement of items.
  - *Delivery Area:* The designated location for delivering completed dishes.
- **Time System:** Captures the temporal dynamics that influence gameplay. It comprises:
  - *Game Clock:* The global time indicator for the game session.
  - *Order Deadlines:* Time constraints associated with each order.
  - *Cooking Timers:* Timers used to manage and monitor cooking processes.

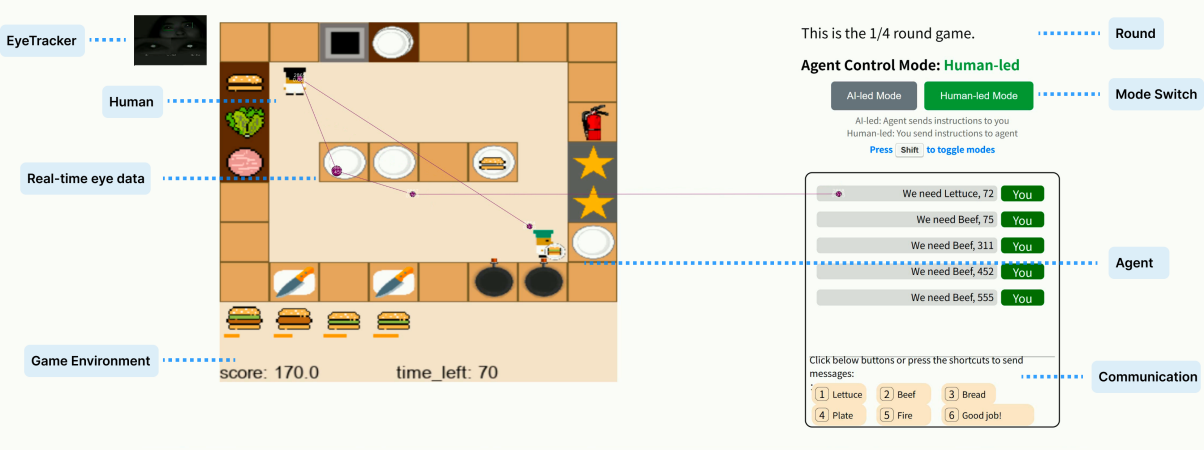

Figure 3: Screenshot of game playing with eye tracking on the upper left side.

4 Methodology

4.1 Participants

Participants were recruited through social media posts and on-campus advertisements. A total of 27 adult participants aged between 18 and 35 took part in the study. No restrictions were placed on prior gaming experience or familiarity with similar games.

Each participant completed a single experimental session that lasted approximately 30–40 minutes. Participants reported a wide range of video game experience and different collision types. All participants were naïve to the specific research objectives prior to participating.

4.2 Ethics

This study adheres to ethical research standards involving human participants. Participants were provided with an informed consent form detailing the study's purpose, their right to withdraw at any time, and how their data would be anonymized and stored.

4.3 General Experiment Setup

We adopt a mixed experimental design, combining both within-subject and between-subject elements. The experiment is conducted on a Windows-based workstation equipped with the Gazepoint GP3 eye-tracker, recording data at 60 Hz. The game interface is displayed on a 24-inch monitor, and participants are seated approximately 60 cm from the screen. Eye-tracking calibration is performed prior to the gameplay. A custom software layer logs all in-game actions, button interactions, and eye-tracking data.

Except for the initial trial round, participants will engage in four rounds of gameplay, structured into two distinct experimental phases:

Baseline Phase (Pre-task) Prior to the main experiment, participants complete a brief low-load visual search task (60 seconds) to establish individualized baseline metrics for pupil dilation. This task is a simple visual engagement with minimal cognitive demand, enabling reliable comparison against task-induced cognitive load during gameplay.

Phase 1 (Rounds 1 and 2) adopts a within-subject design, where each participant experiences both Human-led. The order of mode presentation will be counterbalanced to mitigate ordering effects. This phase aims to observe participants' initial preferences and perceived cognitive load under both conditions. After each round, participants will self-report their perceived cognitive workload using the NASA-TLX questionnaire. Based on these subjective ratings, participants will be classified into either a high-load or low-load group via median split.

Phase 2 (Rounds 3 and 4) adopts a between-subject manipulation based on the cognitive load groupings derived from Phase 1. In Round 3 (the matched condition), participants are assigned a collaboration mode that aligns with their inferred preference (e.g., high-load participants receive AI-led mode). In Round 4 (the mismatched condition), the opposite mode is assigned. This design allows me to assess how alignment or misalignment between cognitive load and collaboration mode influences task performance, user experience, and perceived effort.

4.4 Data Collection

To answer the research questions, data is collected across the following dimensions:

- **Behavioral Data:**
  - Mode-switch frequency and timing (Rounds 1–2)
  - In-game task performance, measured as Completed Task Points (CTP)
  - Collaboration mode selected
- **Physiological Data (via Gazepoint GP3):**

Table 1: Experiment Procedure and Data Collection Overview

Stage	**Round**	**Button Status**	**Collaboration Mode**	**Data Collected**
Baseline	N/A	N/A	N/A (simple visual task)	Pupil dilation (baseline)
Practice	Trial	Enabled	Free collaboration	N/A
Phase 1	Round 1	Enabled	Human-led (default)	NASA-TLX, pupil dilation, mode switch behavior
	Round 2	Enabled	Human-led (default)	NASA-TLX, pupil dilation, mode switch behavior
Phase 2	Round 3	Fake	AI-led	NASA-TLX, pupil dilation, mode switch behavior, task performance
	Round 4	Fake	Human-led	NASA-TLX, pupil dilation, mode switch behavior, task performance

  - Pupil dilation (left eye, right eye)
  - Validity (LPV, RPV)
  - Gaze Coordinates

- **Subjective Measures:**
  - NASA-TLX after first 2 rounds to assess perceived cognitive load
  - Open-ended post-task interviews (e.g., ”Why did you switch modes?”)
  - Observation notes from the experiment

The expected data analysis methods include:

- **RQ1:** To explore the relationship between cognitive load and collaboration mode preference, a correlation analysis was conducted using chi-square tests, with cognitive load discretized into four levels (Low: 0–3, Medium Low: 3–5, Medium High: 5–7, High: 7–10) and collaboration mode preference treated as a binary variable (AI-led vs. Human-led). Both objective cognitive load (pupil dilation) and subjective cognitive load (NASA-TLX ratings) were analyzed separately to capture distinct aspects of cognitive processing..
- **RQ2:** To examine the impact of cognitive load on overall Human–AI team performance, we modeled the relationship between cognitive load and task scores using both linear and quadratic regression. Subjective and objective cognitive load measures were included as predictors, and total performance score was the dependent variable. Quadratic models were used to test for inverted U-shaped relationships, consistent with the Yerkes–Dodson law, which posits optimal performance at moderate levels of arousal or cognitive load.

– **RQ3:** To investigate factors underlying individual performance differences across collaboration modes, we analyzed within-subject variability. First, intraclass correlation coefficients (ICC) were calculated to assess the proportion of variance in cognitive load attributable to between-person versus within-person differences. Next, paired-samples t-tests and matched-versus-unmatched condition comparisons were conducted to examine whether collaboration mode, cognitive load–strategy alignment, or their interaction contributed to performance differences at the round level. This approach allows us to determine whether within-person fluctuations in cognitive load and mode selection are predictive of performance variation beyond stable individual traits.

5 Result

5.1 Data Preparation

For each participant, we collected data from two sources: the eye-tracking log and the game log. The eye-tracking was sampled at 60Hz(60 data points per second), while the game log used step-based time units, with approximately 4 steps per second. First, we cleaned the eye-tracking data by excluding invalid samples: those where both eyes were invalid (LPV = 0 or RPV = 0, indicating the eye tracker could not detect the pupil) or where the gaze coordinates were (0,0), indicating tracking failure.

After cleaning, we aggregated the valid eye-tracking data to match the game log, ensuring that each game step had corresponding eye-tracking information. This allowed us to create a holistic dataframe combining the eye-tracking and game log data for subsequent analysis.

Then objective cognitive load is measured via pupil diameter (left and right eyes). For each participant, the average pupil diameter during a resting baseline phase is first calculated. Momentary cognitive load is then computed using the baseline-relative percentage change formula:

$$\text{Cognitive Load (\%)} = \frac{\text{Current Pupil Diameter} - \text{Baseline}}{\text{Baseline}} \times 100$$

The values from both eyes are averaged to obtain the combined cognitive load, which is then linearly mapped to a 0–10 scale, allowing direct comparison with subjective cognitive load on the same dimension.

As a result, we analyzed data from 27 participants across 4 rounds, with each round containing 800 steps. After excluding invalid data, the dataset contained 79,600 total records for analysis.

5.2 Dispersion of Core Variables

To characterize the variability and distributional properties of the core study variables prior to inferential analysis, we first examined the descriptive statistics of both objective and subjective cognitive load measures.

Across all participants and valid observations, Subjective cognitive load ratings had an average value of 5.66, with a median of 6, indicating that participants generally experienced a moderate level of cognitive demand during the task. The standard deviation of 1.87 suggests noticeable variability in how demanding participants perceived the task to be, with ratings spanning the full range from 1 to 10.

Objective cognitive load showed a slightly lower average value of 4.49 and a median of 4.43. Its smaller standard deviation (1.14) indicates less variability compared to subjective ratings, suggesting that physiological responses were more consistent across observations, despite also covering the full 0–10 scale.

Table 2: Descriptive Statistics of Cognitive Load Measurements

Variable	**Mean (M)**	**SD**	**Min**	**Max**	**Median**
Subjective Cognitive Load (1–10)	5.66	1.87	1.00	10.00	6.00
Objective Cognitive Load (0–10)	4.49	1.14	0.00	10.00	4.43

Building on these descriptive results, we next examined the relationship between objective and subjective cognitive load to determine whether they should be treated as distinct indicators. Pearson's correlation coefficient (r) was used to assess the linear association between the two measures.

The analysis revealed a weak but statistically significant negative correlation between combined objective cognitive load and subjective cognitive load ($r = -0.275, p < .001$). The corresponding coefficient of determination was $R^2 \approx 0.076$, indicating that subjective cognitive load accounts for approximately 7.6% of the variance in objective cognitive load. Conversely, more than 92% of the variance in objective cognitive load remains unexplained by subjective ratings.

Importantly, substantial inter-individual variability was observed in the strength and direction of the correlation, with participant-level correlations ranging from approximately $r = -0.30$ to $r = +0.48$. While some participants exhibited positive associations between subjective and objective measures, others showed negative or near-zero relationships. This high degree of heterogeneity suggests that there is no consistent correlation between subjective and objective cognitive load across individuals.

Taken together, these findings indicate that subjective and objective cognitive load capture related but largely distinct aspects of cognitive processing. As a result, they should be treated as separate indicators rather than combined into a single composite measure in subsequent analyses.

5.3 RQ1: How does cognitive load affect people's preference in human-led mode vs AI-led mode?

To operationalize participants' collaboration mode preference, we used switch attempts as a behavioral indicator. In each of the four rounds, participants could click a switch button to toggle between human-led and AI-led modes (Figure 3). Each click was recorded as a Switch_Attempt = 1, capturing moments when participants actively rejected the current mode.

For rounds where can_switch was FALSE (i.e., the mode could not change), actual switch behavior could not be observed. To maintain consistency across conditions, we approximated hypothetical switch behavior by assigning the alternative collaboration mode at the subsequent time step. Finally, we computed the proportion of switch attempts that led to a transition to AI-led mode, which serves as an empirical measure of AI-led preference. This behavioral metric complements self-reported ratings by capturing participants' active choices.

To examine whether cognitive load was associated with collaboration mode preference, we employed the chi-square test of independence (χ^2). This test is appropriate in the present context for three reasons. First, both variables are categorical: cognitive load was discretized into four levels (Low: 0–3, Medium Low: 3–5, Medium High: 5–7, High: 7–10), and mode preference was binary (AI-led vs. human-led). Second, the dataset contains a large number of observations, satisfying the chi-square test's assumption regarding expected cell frequencies. Third, the chi-square test directly evaluates whether the observed distribution of preferences differs from what would be expected if cognitive load and preference were independent.

Subjective Cognitive Load

Subjective cognitive load showed a significant relationship with ideal mode preference ($\chi^2 = 2660.259$, $p < 0.001$). Participants with low cognitive load (0-3) showed a 32.5% AI-led preference, while those with high cognitive load (7-10) showed only a 25.0% AI-led preference.

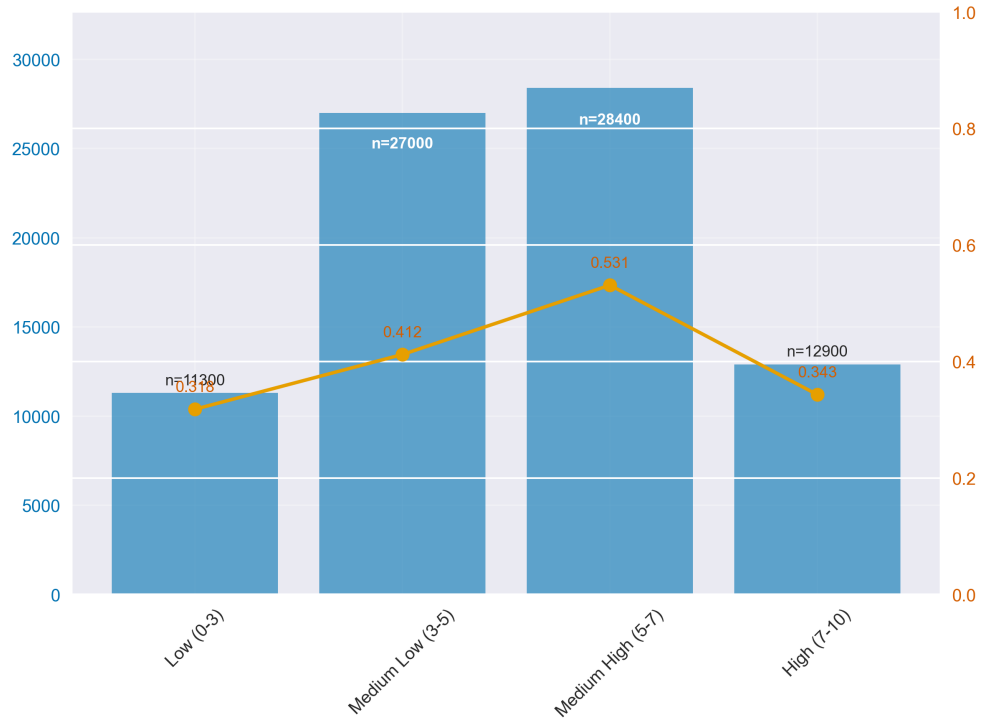

Figure 4: AI-led preference across different levels of subjective cognitive load.

Objective Cognitive Load

Objective cognitive load also demonstrated a significant relationship with ideal mode preference ($\chi^2 = 1653.819$, $p < 0.001$). The pattern showed a clear negative correlation: low objective CL (0-3) had a 55.7% AI-led preference, while high objective CL (7-10) had only an 18.9% AI-led preference.

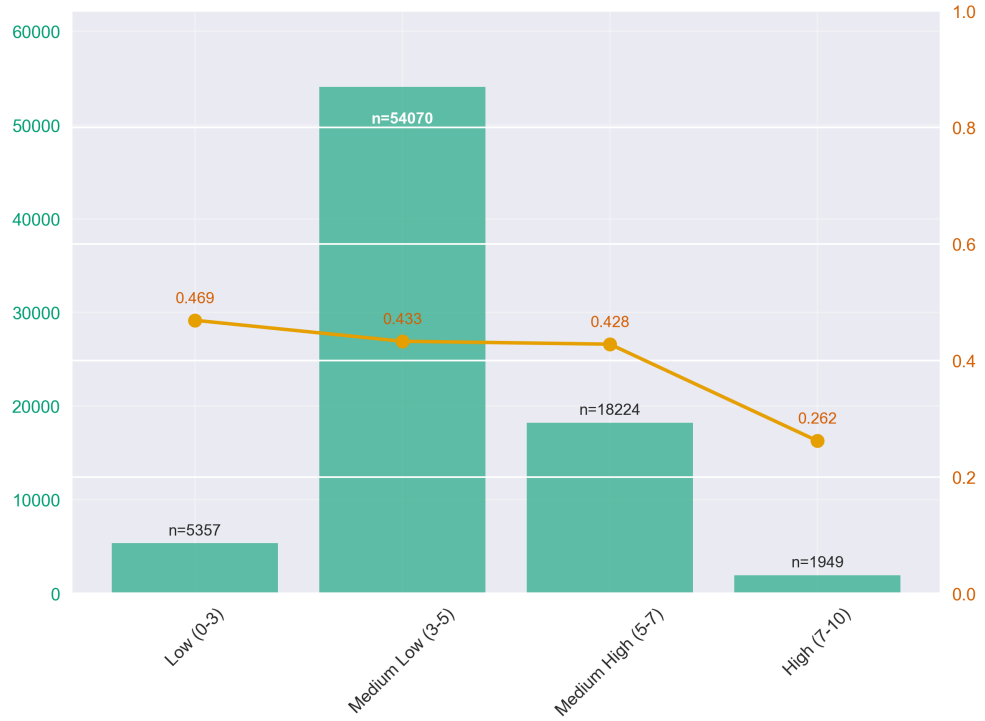

Figure 5: AI-led preference across different levels of objective cognitive load.

The results reveal a non-linear relationship between cognitive load and AI-led preference that aligns with the concept of an optimal cognitive load zone. For both subjective and objective measures, AI-led preference remains relatively stable within the medium cognitive load range (3–7). This suggests that when participants operate under moderate cognitive demand, they are able to effectively collaborate with the AI without feeling either overwhelmed or disengaged.

In contrast, extreme cognitive load levels are associated with more polarized preferences. Under low cognitive load (0–3), participants are more likely to adopt a fully AI-led mode, possibly because the task feels manageable and delegating control to the AI incurs little perceived risk. Conversely, under high cognitive load (7–10), participants show a strong shift toward human-led collaboration, indicating a more conservative strategy in which control is retained by the human when cognitive resources are strained.

Notably, this pattern is substantially stronger for objective cognitive load than for subjective cognitive load, suggesting that physiological strain may exert a more direct influence on behavioral control decisions than conscious self-assessment alone.

5.4 RQ2: How does cognitive load influence performance in Human–AI teams?

The relationship between cognitive load and total performance score was examined using both subjective and objective measures of cognitive load. For both measures, performance exhibited an inverted U-shaped relationship with cognitive load, consistent with the Yerkes–Dodson law of optimal arousal (Yerkes and Dodson, 1908). However, the characteristics of these relationships differed substantially between subjective and objective measures.

For subjective cognitive load, a linear model yielded a modest positive association with performance ($r = 0.256$, $R^2 = 0.066$). In contrast, a quadratic model substantially improved model fit ($r = 0.478$, $R^2 = 0.228$), representing an increase of 0.162 in explained variance. The resulting inverted U-shaped curve peaked at a subjective cognitive load level of 6.11 on a 0–10 scale, indicating that performance was highest at moderate-to-high levels of perceived cognitive load. This pattern suggests that both insufficient and excessive subjective cognitive load are associated with reduced performance, whereas optimal outcomes emerge at intermediate levels (Figure 6).

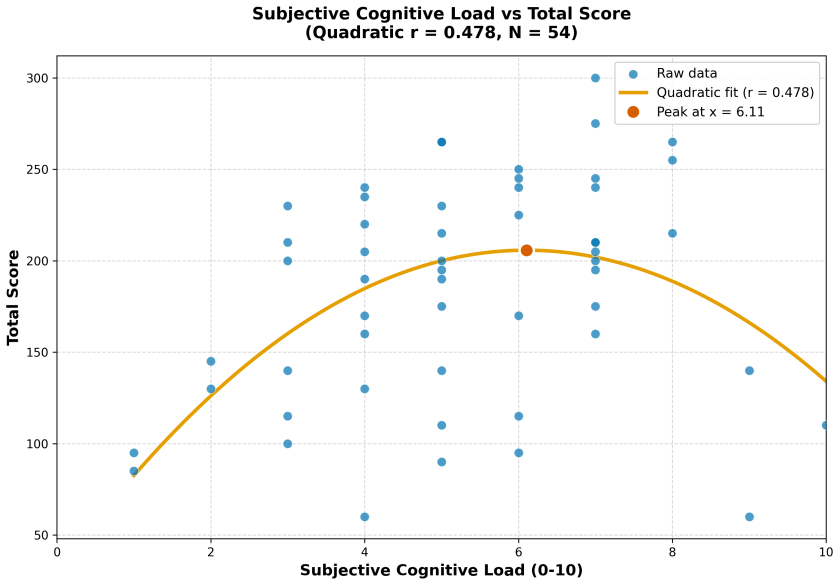

Figure 6: Relationship between subjective cognitive load and performance.

Objective cognitive load demonstrated a similar inverted U-shaped relationship with total score, though with a distinct profile. While the linear association was weak, the quadratic model revealed a clearer pattern, improving model fit to $R^2 = 0.114$ ($r = 0.337$). Notably, the peak of the inverted U-shaped curve occurred at an objective cognitive load level of 3.14 on the same 0–10 scale, substantially lower than the peak observed for subjective cognitive load (6.11). This finding suggests that optimal performance is achieved at relatively low levels of objectively measured cognitive load, in contrast to the moderate-to-high levels indicated by subjective workload ratings (Figure 7).

Several limitations should be noted when interpreting these results. First, the sample size is relatively small, which may constrain the generalizability of the findings. Second, the distribution of cognitive load values is uneven, with observations concentrated between

Figure 7: Relationship between objective cognitive load and performance.

values of 3 and 7, and relatively few data points at the extremes of both cognitive load scales, particularly in the high-load range.

5.5 RQ3: What conditions lead to better individual performance in Human–AI collaboration?

The results of RQ1 demonstrated that cognitive load influences participants' preferences for collaboration strategies, while RQ2 showed that cognitive load is also associated with performance differences at the between-subject level. Building on these findings, RQ3 shifts the focus to the within-subject level, asking what factors account for individual performance differences across collaboration modes.

In particular, we observed substantial within-person variability in performance between Rounds 3 and 4, which corresponded to Human-led and AI-led collaboration modes, respectively. Performance differences across these two rounds ranged from −55 to +85 points, indicating pronounced individual differences that could not be explained solely by group-level effects.

We first examined whether cognitive load varied substantially within individuals across rounds, which would justify treating it as a state-level predictor of performance. To this end, an intraclass correlation coefficient (ICC) analysis was conducted. The ICC quantifies the proportion of total variance attributable to between-person differences relative to within-person variability across repeated measurements. An ICC value close to 1 indicates that most variance arises from stable between-person differences (i.e., trait-like characteristics), whereas an ICC value close to 0 suggests that variance is primarily driven by within-person fluctuations across rounds (i.e., state-like variation).

The analysis included data from 27 participants, each completing 4 experimental rounds. For each participant, mean cognitive load was first computed at the round level by averaging all timestep-level measurements within each round, yielding 4 round-level mean values per participant. Based on these values, the ICC was calculated as:

$$\text{ICC} = \frac{\sigma^2_{\text{between}}}{\sigma^2_{\text{between}} + \sigma^2_{\text{within}}}, \tag{1}$$

where $\sigma^2_{\text{between}}$ represents the between-person variance, operationalized as the variance of participant-level means averaged across the 4 rounds, and σ^2_{within} represents the within-person variance, defined as the average variance of round-level means within each participant.

The results indicated a very high ICC for objective cognitive load ($\text{ICC} = 0.9701$), suggesting that 97.0% of the variance was attributable to between-person differences, with only 3.0% attributable to within-person variability across rounds. For subjective cognitive load, the ICC was also high ($\text{ICC} = 0.7521$), indicating that 75.2% of the variance reflected between-person differences and 24.8% reflected within-person variability.

Taken together, these results suggest that both objective and subjective cognitive load function predominantly as stable, trait-like individual characteristics in this task, rather than as highly fluctuating state variables across rounds. As a result, cognitive load alone is unlikely to account for the large within-person performance differences observed between collaboration modes.

Next, we examined whether collaboration mode exerted a main effect on performance. A paired-samples t-test comparing Human-led and AI-led modes revealed a mean performance difference of +8.33 points, which was not statistically significant ($p > 0.05$). This indicates that collaboration mode alone does not have a significant main effect on performance.

Finally, we examined whether the match between cognitive load and collaboration strategy contributed to performance differences. Performance was compared between matched and unmatched conditions for both objective and subjective cognitive load. For objective cognitive load, no significant performance advantage was observed for matched relative to unmatched conditions. For subjective cognitive load, matched participants showed numerically higher performance; however, the small sample size limits the robustness and generalizability of this effect. Overall, these findings suggest that cognitive load–strategy alignment is not a primary driver of within-person performance differences in this task.

6 Discussion

6.1 Explorative Analysis

In this work, we set out to explore human factors, particularly cognitive load, as a window into human–AI teaming. By examining both cognitive load and user preferences toward collaboration modes, our findings revealed not only patterns in human mental load but also meaningful differences in how individuals approach collaboration. While cognitive load was our primary focus, this study is intentionally exploratory. As such, we also would like to discuss additional factors that may influence cognitive load during interaction.

One such factor is collaboration style. In the post-experiment questionnaire, participants were asked whether they perceived themselves as individual contributors (IC) or team

Table 3: Ideal cognitive load range (3-7) for individual contributors (IC) and team players (TP).

Comparison	IC (%)	TP (%)	Difference (pp)	t-test	p	Cohen's d
Overall	84.81	94.54	9.73	$t(25) = -0.623$	0.539	0.244

players (TP). This self-reported classification provided useful context for interpreting behavioral and cognitive differences observed during gameplay.

We conducted an exploratory, descriptive analysis to examine whether differences in cognitive load regulation between team players and individual contributors were also reflected in task performance. Given the small sample size, this analysis was not intended as a confirmatory test, but rather to identify potential trends for future investigation.

27 participants were classified as individual contributors (IC; n = 11) or team players (TP; n = 16) based on their self-reported responses to a post-experiment questionnaire. Given the nested structure of the data, with multiple timesteps per participant, cognitive load measures were aggregated at the participant level to satisfy the independence assumption. Each participant therefore contributed a single data point representing their average proportion of time spent within the ideal cognitive load range.

As shown in Table 9, TP participants spent a higher proportion of time within the ideal cognitive load range (94.54%) than IC participants (84.81%), corresponding to a difference of 9.73 percentage points. Although the between-group t-test did not reach statistical significance (p = 0.539), the observed effect size (d = 0.244) suggests a small-to-moderate effect in favor of team players.

One possible explanation relates to perceived control. Individual contributors may place greater emphasis on personal ownership and autonomy during task execution. When the AI assumes a leading role, this perceived loss of control may introduce friction, potentially impairing cognitive load regulation. While this interpretation remains speculative given the exploratory nature of the analysis and limited sample size, the consistent direction of effects suggests a meaningful trend that merits further investigation.

Qualitative observations further support this interpretation. Player 7, who identified as an individual contributor, experienced noticeable difficulties during collaboration. Although the game was explicitly framed as a cooperative task, this participant tended to work independently from the AI. When the AI completed the serving step after the human had plated the order, the participant remarked, "AI steals my order," indicating friction and misalignment in collaborative expectations.

In contrast, participants who identified as team players demonstrated smoother and more adaptive collaboration with the AI. For example, Players 5 and 13 coordinated seamlessly with the AI despite holding opposite preferences for collaboration mode (human-led vs. AI-led). Both described the AI as a teammate and reported actively adapting their strategies to optimize joint performance. This suggests that in human–AI teaming, individuals may transfer prior experience and mental models from human–human collaboration, which in turn supports more effective cognitive load regulation.

6.2 Limitation and future work

This study has several limitations that point to promising directions for future work. First, isolating cognitive load from other influencing factors in human–AI collaboration

remains inherently challenging. In this work, cognitive load serves as an interpretive lens for understanding human behavior and performance rather than as a standalone construct. In complex collaborative scenarios, multiple factors, such as perceived control, task ownership, communication dynamics, and prior experience with AI—are likely to jointly influence observed cognitive load. As a result, it is difficult to unravel the precise mechanisms through which the cognitive load emerges. Future work could employ more fine-grained experimental manipulations or multimodal measurements (e.g., physiological signals or behavioral markers) to better disentangle these contributing factors.

A more classical limitation of this study is the relatively small sample size. In addition, as mentioned, participants were adults aged between 18 and 30, which limits the generalizability of the findings. Age-related familiarity with AI may systematically influence collaboration strategy choice and cognitive load, with younger users may more readily adopting AI-led collaboration and older users may experience higher coordination demands. Future studies with larger and more diverse samples are therefore necessary to assess the robustness and generalizability of the findings.

Finally, the primary goal of this study was to explore how cognitive load, as a key human factor, is related to interaction dynamics and performance in human–AI teams. Consequently, the study does not aim to characterize extreme states such as cognitive overload or underload, but rather focuses on how variations in cognitive load within a typical interaction range shape collaboration outcomes. As such, the findings should be interpreted as exploratory rather than confirmatory. Future work could explicitly investigate boundary conditions, including overload and underload scenarios, to better understand how human–AI collaboration breaks down or adapts under more demanding or less engaging conditions.

7 Conclusion

This work investigated cognitive load as a key human factor in human–AI collaboration using an Overcooked-inspired cooperative game equipped with real-time eye tracking. Participants collaborated with an AI teammate under different leadership settings, while cognitive load was continuously estimated and complemented by subjective self-reports. This experimental setup provides a window into how humans adapt their interaction strategies during real-time collaboration with AI agents, and how these adaptations may relate to task performance.

Subjectively, participants who experienced cognitive overload or underload tended to prefer human-led collaboration, potentially reflecting a desire to retain control or reduce coordination uncertainty. Objectively, lower cognitive load was associated with increased use of AI-led collaboration, suggesting that users may be more willing to delegate control to AI when cognitive resources are available. Although no significant impact on task performance was observed, which is likely due to the limited sample size, these findings highlight the role of cognitive load in shaping interaction dynamics and user preferences. More broadly, the results suggest design opportunities for adaptive AI systems that respond to users' cognitive states rather than relying solely on static interaction modes.

Acknowledgements

This thesis marks one of the most challenging projects I have undertaken during my studies. I would like to express my sincere gratitude to my supervisors, Max van Duijn provided invaluable guidance on the overall design and direction of my thesis, while Michiel van der Meer devoted significant time and effort to discussions and close supervision throughout the research process. Without their support, insight, and patience, this work would not have been possible; I would also like to thank my friends and family for their constant encouragement. In particular, I am grateful to those who participated in my research: their genuine interest in the topic repeatedly inspired me to delve deeper and stay committed to this work; Finally, I want to thank myself. Studying in the Netherlands for two and a half years has been a journey filled with uncertainty, courage, and resilience. This experience has led me to a landscape I never expected, one that I embraced with curiosity and bravery.

References

Akata, Z., Balliet, D., de Rijke, M., Dignum, F., Dignum, V., Eiben, G., Fokkens, A., Grossi, D., Hindriks, K., Hoos, H., Hung, H., Jonker, C., Monz, C., Neerincx, M., Oliehoek, F., Prakken, H., Schlobach, S., van der Gaag, L. C., van Harmelen, F., van Hoof, H., van Riemsdijk, B., van Wynsberghe, A., Verbrugge, R., Verheij, B., Vossen, P., and Welling, M. (2020). A research agenda for hybrid intelligence: Augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. *Computer*, 53(8):18–28.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M. T., and Weld, D. (2021). Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–16.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, 91(2):276–292.

Berberian, B., Sarrazin, J.-C., Le Blaye, P., and Haggard, P. (2012). Automation technology and sense of agency: A window on human–machine interaction. *Human Factors*, 54(3):386–397.

Brynjolfsson, E., Rock, D., and Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. *The Economics of Artificial Intelligence*.

Byrne, M. D. and Parasuraman, R. (1996). Psychophysiology and adaptive automation. *Biological Psychology*, 42(3):249–268.

Carroll, M., Shah, R., Ho, M. K., Batra, D., Abbeel, P., and Dragan, A. D. (2019a). On the utility of learning about humans for human-ai coordination. *Advances in Neural Information Processing Systems (NeurIPS)*, 32.

Carroll, M., Shah, R., Ho, M. K., Griffiths, T., Seshia, S., Abbeel, P., and Dragan, A. D. (2019b). On the utility of learning about humans for human–ai coordination. *Advances in Neural Information Processing Systems*, 32.

Desmond, P. A. and Hancock, P. A. (1998). Active and passive fatigue states. *Human Factors*, 40(3):515–520.

Fitts, P. M. (1951). Human engineering for an effective air-navigation and traffic-control system. *Journal of the Institute of Aeronautical Sciences*, 18(3):144–152.

Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2022). Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation. *Information Systems Research*, 33(2):678–696.

Gopher, D. and Kimchi, R. (1989). Engineering psychology. *Annual Review of Psychology*, 40:431–455.

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y. C., de Visser, E. J., and Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human–robot interaction. *Human Factors*, 53(5):517–527.

Hancock, P. A. and Caird, J. K. (1993). Experimental evaluation of a model of mental workload. *Human Factors*, 35(3):413–429.

Hemmer, P., Schemmer, M., Vössing, M., and Kühl, N. (2021). Human-ai complementarity in hybrid intelligence systems: A structured literature review. In *Proceedings of the Pacific Asia Conference on Information Systems (PACIS)*, pages 1–14.

Ho, M. K., Ermon, S., and Dragan, A. D. (2019). Human-aware planning in human-ai collaboration. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*, pages 1500–1508.

Hockey, G. R. J., Briner, R. B., Tattersall, A. J., and Wiethoff, M. (1989). Assessing the impact of workload on operator stress: The role of control. *Human Factors*, 31(6):659–673.

Holmes, W., Bialik, M., and Fadel, C. (2019). *Artificial Intelligence in Education: Promises and Implications for Teaching and Learning*. Center for Curriculum Redesign, Boston, MA, USA.

Hu, R., Wang, X., Zhang, W., and Chen, Y. (2022). Adapting to human teammates: Theory-of-mind-based human-ai co-adaptation. In *Proceedings of the 21st International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Huang, S., Chen, T., Zhang, F., Sun, J., Li, X., Cheng, Y., and Wang, W. (2024). A survey on large language model based autonomous agents. *arXiv preprint*.

Johnson, M., Bradshaw, J. M., Feltovich, P. J., Jonker, C. M., van Riemsdijk, M. B., and Sierhuis, M. (2014). Coactive design: Designing support for interdependence in joint activity. *Journal of Human-Robot Interaction*, 3(1):43–69.

Krejtz, K., Duchowski, A. T., Niedzielska, A., Biele, C., and Krejtz, I. (2018). Eye tracking cognitive load using pupil diameter and microsaccades with fixed gaze. *PLOS ONE*, 13(9):e0203629.

Laeng, B., Sirois, S., and Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, 7(1):18–27.

Lee, J., Davari, H., Singh, J., and Pandhare, V. (2018). Industrial ai: Applications with sustainable performance. *Journal of Manufacturing Systems*, 48:144–152.

Li, Z., Zhu, H., Lu, Z., Xiao, Z., and Yin, M. (2025). From text to trust: Empowering ai-assisted decision making with adaptive llm-powered analysis. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, Yokohama, Japan. CHI '25.

Parasuraman, R. and Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, 39(2):230–253.

Parasuraman, R., Sheridan, T. B., and Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., and Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *arXiv preprint*.

Recarte, M. A. and Nunes, L. M. (2000). Mental workload while driving: Effects on visual search, discrimination, and decision making. *Journal of Experimental Psychology: Applied*, 6(4):345–362.

Rubio, S., Díaz, E., Martin, J., and Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of swat, nasa-tlx, and workload profile methods. *Applied Psychology*, 53(1):61–86.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., Maier, R., Merz, A. B., Oeste-Reiß, S., and Randrup, N. (2020). Machines as teammates: A research agenda on ai in team collaboration. *Information & Management*, 57(2):103174.

Sevcenko, N., Ninaus, M., Wortha, F., Moeller, K., and Gerjets, P. (2021). Measuring cognitive load using in-game metrics of a serious simulation game. *Frontiers in Psychology*, 12:572437.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Sweller, J., van Merriënboer, J. J. G., and Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, 10(3):251–296.

Topol, E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56.

Vaccaro, M., Almaatouq, A., and Malone, T. (2024). When combinations of humans and ai are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(12):2293–2303.

van der Wel, P. and van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, 25:2005–2015.

Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., and Anandkumar, A. (2023). Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors*, 50(3):449–455.

Wilson, J. R. and Rajan, J. (1995). Human factors considerations in reliability engineering. *Reliability Engineering & System Safety*, 49(3):335–346.

Wiltgen, B., Hemmer, P., and Kühl, N. (2024). Why stronger models do not automatically improve human–ai coordination. *arXiv preprint*.

Wu, Y., Zhang, S., Wang, X., Li, C., and Chen, Y. (2021). Agent modeling and human-ai co-adaptation in repeated interactions. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems (AAMAS)*.

Yerkes, R. M. and Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology*, 18(5):459–482.

Young, M. S. and Stanton, N. A. (2002). Mental workload: Theory, measurement, and application. *Human Factors*, 44(3):332–340.

Yuksel, B. F., Oleson, K. B., Harrison, L. T., Peck, E. M., Afergan, D., Chang, R., and Jacob, R. J. K. (2016). Learn piano with bach: An adaptive learning interface that adjusts task difficulty based on brain state. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 5372–5384.

Zhang, S., Wang, X., Zhang, W., Li, C., Song, J., Li, T., Qiu, L., Cao, X., Cai, X., Yao, W., and Wen, Y. (2025). Leveraging dual process theory in language agent framework for real-time simultaneous human-ai collaboration.

Appendix A: Main Table Structure

The aggregated dataset `combined_aggregated_data_filled_pupil_with_0to10.csv` contains timestep-level observations combining game log data, eye-tracking measurements, and questionnaire responses.

- **Total observations**: 79,600 rows (timestep-level data points)
- **Total variables**: 74 columns
- **Data structure**: Each row represents one timestep for one participant in one round
- **Rounds included**: 0 (practice), 1, 2, 3 4

Variable List by Category

Identification Variables (3)

- `participant_id`: Participant identifier
- `round_id`: Round number (0, 1, 2, 3, 4)
- `timestep`: Timestep index within the round

Game Performance Variables (4)

- `score`: Score gained in this timestep
- `total_score`: Cumulative score up to this timestep
- `n_orders`: Number of active orders
- `n_objects`: Number of objects in the game state

Communication Variables (4)

- `n_messages`: Total number of messages in this timestep
- `messages`: List of messages (JSON format)
- `n_human_message`: Number of messages in Human-led mode
- `n_ai_message`: Number of messages in AI-led mode

Action Variables (5)

- `human_action`: Human player's action code
- `ai_action`: AI agent's action code
- `has_assigned_tasks`: Boolean indicating if tasks were assigned
- `n_assigned_tasks`: Number of assigned tasks
- `assigned_tasks`: List of assigned tasks (JSON format)

Experimental Condition Variables (4)

- `team_player`: Player type (0=Individual Contributor, 1=Team Player)
- `Current_Mode`: Collaboration mode (0=Human-led, 1=AI-led)

- `can_switch`: Boolean indicating if mode switching is allowed
- `Switch_Attempt`: Boolean indicating if a mode switch was attempted

Subjective Measure Variables (1)

- `subjective_cognitive_load`: Self-reported cognitive load from questionnaire

Eye-tracking Variables - Left Eye (15 variables)

- `avg_lpcx`, `std_lpcx`, `count_lpcx`: Left pupil center X (mean, std, count)
- `avg_lpcy`, `std_lpcy`, `count_lpcy`: Left pupil center Y (mean, std, count)
- `avg_lpd`, `std_lpd`, `count_lpd`: Left pupil diameter (mean, std, count)
- `avg_lps`, `std_lps`, `count_lps`: Left pupil size (mean, std, count)
- `avg_lpv`, `std_lpv`, `count_lpv`: Left pupil velocity (mean, std, count)

Eye-tracking Variables - Right Eye (15 variables)

- `avg_rpcx`, `std_rpcx`, `count_rpcx`: Right pupil center X (mean, std, count)
- `avg_rpcy`, `std_rpcy`, `count_rpcy`: Right pupil center Y (mean, std, count)
- `avg_rpd`, `std_rpd`, `count_rpd`: Right pupil diameter (mean, std, count)
- `avg_rps`, `std_rps`, `count_rps`: Right pupil size (mean, std, count)
- `avg_rpv`, `std_rpv`, `count_rpv`: Right pupil velocity (mean, std, count)

Eye-tracking Variables - Saccades (6 variables)

- `avg_saccade_mag`, `std_saccade_mag`, `count_saccade_mag`: Saccade magnitude
- `avg_saccade_dir`, `std_saccade_dir`, `count_saccade_dir`: Saccade direction

Eye-tracking Variables - Blinks (9 variables)

- `avg_bkid`, `std_bkid`, `count_bkid`: Blink ID
- `avg_bkdur`, `std_bkdur`, `count_bkdur`: Blink duration
- `avg_bkpmin`, `std_bkpmin`, `count_bkpmin`: Blinks per minute

Eye-tracking Metadata Variables (5)

- `total_samples`: Total eye-tracking samples in the window
- `window_start`: Window start timestamp
- `window_end`: Window end timestamp
- `real_timestamp`: Real timestamp
- `has_eye_data`: Boolean indicating if eye-tracking data is available

Cognitive Load Variables - 0-10 Scale (3 variables)

- `cognitive_load_lpd_0to10`: Left pupil diameter on 0-10 scale
- `cognitive_load_rpd_0to10`: Right pupil diameter on 0-10 scale
- `cognitive_load_combined_0to10`: Combined pupil diameter on 0-10 scale

Sample Data (First 5 Rows, Selected Columns)

Participant	Round	Timestep	Total Score	Messages	Mode	Team Player	Subj. CL	Combined CL
1	9	0	0.0	0	Human-led	IC	9	3.77
1	9	1	0.0	0	Human-led	IC	9	3.59
1	9	2	0.0	0	Human-led	IC	9	3.35
1	9	3	0.0	0	Human-led	IC	9	3.52
1	9	4	0.0	0	Human-led	IC	9	3.39

Table 4: Sample data from the aggregated dataset (showing key columns only). The complete dataset contains 79,600 rows and 74 columns.