Who Did What? Designing Avatars for Explainable Multi-Agent Systems in Knowledge Work

Simon Rapp
human-centered systems lab (h-lab)
Karlsruhe Insititute of Technology
Karlsruhe, Baden-Württemberg, Germany
simon.rapp@kit.edu

Martin Feick
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
martin.feick@kit.edu

Marcus Jainta
EnBW AG
Karlsruhe, Germany
m.jainta@enbw.com

Alexander Maedche
human-centered systems lab (h-lab)
Karlsruhe Institute of Technology (KIT)
Karlsruhe, Germany
alexander.maedche@kit.edu

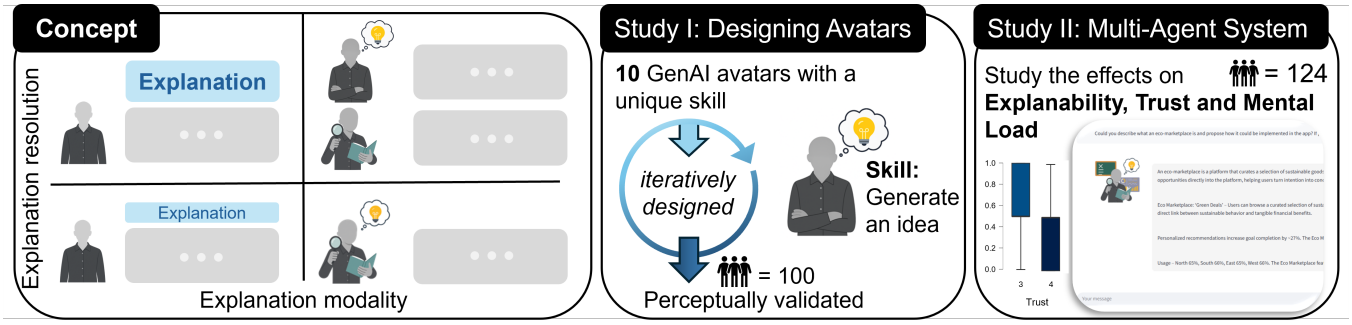

Figure 1: Left: Explanation concepts varying in explanation modality (text vs. avatars) and explanation resolution (low vs. high). Middle: Study I — iterative design of 10 avatars representing a unique skill, perceptually validated with 100 participants. Right: Study II — mixed-methods evaluation with a working multi-agent system (online experiment: N = 124; follow-up interviews with a randomly selected subset of participants, N = 20), examining the effects on explainability, trust and mental load.

Abstract

Knowledge workers increasingly rely on multi-agent systems to solve complex problems. While these systems offer valuable support, they often obscure which agents contributed to a response, leading to a lack of transparency that may result in errors and reduced trust. To address this, we propose avatars that make agents' expertise and contributions transparent. We iteratively co-designed avatars representing distinct expertise areas and validated them in an experiment (N=100). Building on this, we developed four multi-agent prototypes varying in explanation modality (text vs. avatars) and resolution (low vs. high). We then conducted a mixed-methods evaluation with an online experiment (N=124) and follow-up interviews (N=20). Qualitative results suggest that avatars foster clearer mental models, improve perceived explainability, and support users' trust calibration without increasing cognitive load, although no significant quantitative differences were found. Our research contributes validated avatar designs, insights into explanation strategies, and design implications for explainable multi-agent systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**.

Keywords

Multi-Agent System, Explainability, Avatars

ACM Reference Format:
Simon Rapp, Martin Feick, Marcus Jainta, and Alexander Maedche. 2026. Who Did What? Designing Avatars for Explainable Multi-Agent Systems in Knowledge Work. In *Designing Interactive Systems Conference (DIS '26), June 13–17, 2026, Singapore, Singapore.* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/3800645.3812981

1 Introduction

Imagine Tina, a delivery manager who is heading an organization that is running many important customer projects simultaneously. To effectively handle the demands of her knowledge work job and to manage her workload efficiently, she uses specialized Conversational Agents (CAs), e.g., powered by Large Language Models (LLMs). These specialized CA are fine-tuned to excel in specific tasks, for example, helping Tina with computing the most

This work is licensed under a Creative Commons Attribution 4.0 International License.
DIS '26, Singapore, Singapore
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2563-0/26/06
https://doi.org/10.1145/3800645.3812981

recent quarterly sales figures. However, single CAs struggle with complex, multi-faceted problems [3], e.g., Tina does not only need to summarize the sales figures, but also needs to create a report and charts for the customer presentation. Here, Multi-Agent Systems (MASs) can provide a solution by dynamically combining several specialized agents that collaboratively address such problems [44]–this is how, for example, Microsoft Copilot can be deployed within organizations [26]. One agent computes the sales figures, while another builds on the output to create the report and charts.

However, current MASs can provide limited explainability to end users. When Tina uses the system, she may be unaware of which agents were involved or what their contributions were, for example, in systems such as Microsoft Copilot [26]. This lack of attribution transparency can hinder users' understanding and appropriate trust calibration [10, 16, 23]. At the same time, users must rely on the MAS output without an easy way to verify how it was produced, which may make it harder to appropriately assess and act on outputs [29, 40]. Ultimately, relying on outputs without insight into agent contributions may increase the risk of errors in knowledge work. To address this, related work proposed displaying detailed textual explanations, e.g, the step-by-step reasoning created by Chain-of-Thoughts (CoTs), to help Tina to understand how the system arrived at its answer [36]. However, findings showed that excessive textual explanations may increase cognitive load and lead to Tina being overwhelmed [43]. Thus, Tina needs a way to quickly verify if the most suitable agents have contributed to the output without adding a high level of complexity.

This is where avatars, i.e., visual representations of agents, come in. They offer a promising approach to enhance explainability of MASs while avoiding additional cognitive load. By visually representing agents, avatars could help users understand which agents were involved and what their contributions to the MASs output were, all without requiring any additional textual explanation [27, 32]. However, little is known about how avatars for specialized CAs should be designed to represent their area of expertise for a knowledge worker like Tina, leading to the question: ***RQ1: How can avatars be designed to effectively and intuitively represent their area of expertise?*** To investigate this, we informed our iterative co-design approach with semiotic theory and designed 11 avatars with and without visual cues (see Figure 3). 10 avatars reflected typical tasks that knowledge workers face, such as summarizing facts or creating new ideas, according to Brachman et al. [4]. In an online within-subjects experiment ($N = 100$), we evaluated whether our avatars effectively conveyed expertise. Our results show that 6 out of 11 designed avatars significantly represent a unique area of expertise.

However, MASs do not only utilize one agent at a time, but make use of multiple agents to solve complex user requests. To represent this key characteristic, we designed two variations (see Figure 5): (1) a merged avatar, combining different expertise cues and providing one output, and (2) multiple single avatars that represent their individual contributions to the MASs response. In the next step, we implemented a fully functional MAS to test how our avatars perform against textual explanations for typical tasks in knowledge work. We followed a mixed-methods approach, combining an online experiment ($N = 124$) with follow-up interviews conducted with a randomly selected subset of participants ($N = 20$). In the online between-subjects experiment, we compared different *explanation modalities* (text vs. avatars) and levels of *explanation resolution* (low vs. high). In the latter, the MAS shows a detailed textual explanation of which specialized agents were used and for what purpose, or provides abbreviated textual explanations specifying only which specialized agents were used (see Figure 5). In the avatar condition, we included the merged avatar and the individual avatars. Overall, we aim to understand ***RQ2: How do explanation modality (text vs. avatars) and explanation resolution (low vs. high) affect users' perceived explainability, trust, and cognitive load.***

While our quantitative results did not reveal significant differences between conditions, including in cognitive load, our qualitative findings from the interviews indicate that distinct avatars next to agent outputs were perceived as helpful in forming accurate mental models, while participants did not report higher cognitive load. We contribute to the emerging stream of human-Artificial Intelligence (AI) interaction by:

(1) Designing and empirically validating a set of avatars for specialized CA for knowledge work.
(2) Empirically evaluating the effects of different types of explanations in a MAS on key variables in knowledge work.
(3) Providing design implications to enhance explainability in future MAS.

2 Related Work

2.1 Human-AI Collaboration in Knowledge Work

Knowledge workers are individuals with advanced expertise, education, or experience, whose roles primarily involve creating, sharing, and applying knowledge [7]. In 2024, three out of four knowledge workers worldwide reported using AI at work [25]. Using AI, such as LLMs, helps knowledge workers like Tina to save time, focus on important work, enhance their creativity, and enjoy their work more [25]. Recently, Brachman et al. [4] investigated typical tasks, where knowledge workers rely on or use LLMs in an enterprise context. They provide an overview of tasks such as the creation of ideas, information seeking, and summarizing text or validating rules. This builds the basis of our investigations.

Today, LLM-based agents can be fine-tuned, and thereby specialized, to achieve high performance in specific knowledge work tasks [3]. These specialized LLM-based agents struggle with complex, multi-faceted problems [3]. Here, MASs provide a solution by dynamically combining several specialized agents that collaboratively address these problems [44]. Building on this idea, recent frameworks such as *AutoGen* [45] demonstrate how multiple LLMs-based agents can adopt complementary roles (e.g., agent to write code or execute code). These frameworks illustrate a shift from single-task execution toward collaborative team processes, where agents contribute diverse perspectives to achieve complex goals [31]. MASs allow the exposition of how different agents contributed to a shared outcome. However, little is known about how users perceive and interact with such systems, especially in the context of collaborative knowledge work [37]. In this regard, prior work in HCI has highlighted the importance of aligning AI-based systems with human-centered design guidelines to support effective interaction, trust, and user understanding [2]. Complementing this, Yun

et al. [46] argue that knowledge workers benefit from transparent collaboration mechanisms and controllable AI behavior to synthesize information. This motivates our focus on explainability in MASs, as user understanding of agent contributions seems essential for effective collaboration.

2.2 Explainability in Multi-Agent Systems

Explainability in MASs refers to also making transparent which agents were involved and what they contributed to the output [10, 16, 23]. In contrast to explainability in single-agent systems, where the focus lies on justifying one agent's reasoning (e.g., by displaying textual explanations like the step-by-step reasoning of a CoT), MASs also require attribution transparency across multiple agents [16, 23, 36]. Prior work on end-user explainability needs regarding AI highlights that users often seek practically useful information to improve collaboration with AI, rather than technical system details [21]. Despite growing interest in explainability for MASs, little research has investigated how knowledge workers can be effectively supported in understanding agent expertise and contributions in MAS.

However, two recent studies suggest promising design directions. Song et al. [37] show that making multiple AI agents visible can shape user perceptions through social influence, indicating that exposing agent contributions may affect how users form attitudes toward MASs. Complementing this, Schelhorn et al. [36] demonstrate that displaying step-by-step reasoning improves users' understanding and certainty in analytic contexts. Together, this points to two complementary modalities for explainability in MASs: (1) non-textual visual differentiation of agents [20, 27, 37], and (2) textual explanations [22, 36]. Explainability in MASs involves a key tension: too little transparency risks opacity and mistrust, while too much detail can overwhelm users. Prior work highlights this trade-off, as developers often require detailed visibility for debugging, whereas end users tend to prefer simplicity and seamless integration [29]. Related research shows that approaches such as progressive disclosure and visual explanation cues can help balance transparency and cognitive load, but excessive detail may still lead to overload [40]. Moreover, increasing the completeness of explanations can improve understanding, while overly simplified explanations may undermine trust and usability [22]. Beyond MASs, transparency has been shown to support users' mental models, emphasizing the need to tailor explanations to user needs [9].

Abdul et al. [1] outline an HCI research agenda for explainable systems, emphasizing the need for intelligible user interfaces that make system behavior understandable and accountable to end users. Our work responds to this call by empirically investigating how explanation modality (text vs. avatars) and resolution (low vs. high) affect knowledge workers' perceived explainability, trust, and cognitive load in MASs. While agent expertise can be conveyed textually, representing it in a non-textual, visual manner remains challenging. We therefore explore avatars as a means to make agent expertise and contributions more salient.

2.3 Avatars for Specialized Agents

Research on visual histories shows that visual elements, like screenshots, enable users to rapidly recognize prior activities, reconstruct mental context, and navigate complex digital environments more effectively than with textual traces alone [18, 34]. Avatars have been widely used as visual representations of CA in virtual environments [20, 32, 41]. Previous work shows that avatars can support recognition of the active agent and help distinguish between multiple agents [20, 27]. Beyond recognition, avatars shape how agents are perceived: by adding a sense of authenticity and social presence, they may foster user satisfaction and trust [11, 12, 19, 32]. Recent work also shows that avatar representations influence user experiences in nuanced ways, including prompting behaviors and perceived human touch [39]. The design of avatars requires careful calibration. For instance, avatars representing agents with human-like skills (e.g., idea generation) are often perceived as more trustworthy when they convey human-likeness without hyper-realistic detail, as this supports social presence while avoiding the Uncanny Valley [28, 32]. Wallkötter et al. [42] show that embodied agents can use social cues to externalize internal states and thereby support explainability. While their work focused on dynamic behavior, the principle suggests that avatars in MASs could also employ specific visual cues (e.g., symbols or props) to signal agent expertise.

Research Gap. Although research on CAs has increasingly focused on avatars [32], these are often designed for general-purpose assistants or social interaction contexts rather than functionally specialized agents such as those tailored for ideation or document summarization. At the same time, prior research has advanced our understanding of explainability in MASs and agent support in knowledge work, yet little is known about how these dimensions intersect. In particular, it remains unclear how different avatars and textual modalities for communicating agent expertise and contribution influence knowledge workers perceived explainability, trust, and cognitive load in MASs.

3 Study I: Designing Avatars with Expertise Cues

In the first study, we set out to design avatars containing visual cues that, without additional training, can be associated with a single area of expertise.

3.1 Co-Design Process

The avatar co-design process was conducted in three iterations in collaboration with external researchers (N = 12) who were not part of the research team. Four participants were from industry and eight were affiliated with our institution. We recruited researchers who use LLM-based systems, as they represent the target user group (knowledge workers) for our study. In the initial iteration, all authors brainstormed and sketched initial avatar ideas. These sketches informed subsequent AI prompts in the next iteration, where we used a generative AI tool (ChatGPT-4) to create initial candidate avatars. Participants were involved in two following iterations to review, discuss, and refine the designs as illustrated in Figure 2, where the orange arrows indicate the co-design iterations. An example AI prompt can be found in the supplementary material. This approach enabled rapid design exploration. By generating avatars with AI rather than using existing libraries or manual refinement, we avoided copyright concerns and ensured that the avatars can be freely adapted and reused in future work, including by the HCI

Figure 2: Iterative co-design process for one example avatar.

community and corporate settings. Moreover, this approach helped overcome limitations in our own artistic skills, allowing us to produce high-quality avatar designs.

Design Iterations. In total, 11 AI avatars were designed: 10 were intentionally aligned with distinct areas of expertise, while one was created as a control avatar without any intended expertise. We designed human-like avatars to reflect human-like abilities, such as generating ideas [32]. To improve natural interaction with symbolic elements (e.g., pointing to a chalkboard), we also agreed to depict the avatars from the waist up, allowing them to act on objects in ways consistent with human gestures. The avatars, especially their visual cues, are informed by semiotic theory, which conceptualizes signs as carriers of meaning that link a perceivable form (signifier) with an interpreted concept (signified) [8, 30]. Through two co-design iterations, 12 participating researchers evaluated the 11 avatars in an online survey. Each researcher was sequentially shown all 11 avatars in randomized order and rated the extent to which each avatar embodied expertise across ten predefined areas on a 5-point Likert scale. In addition, researchers provided open-ended feedback and design suggestions for each avatar. Based on this feedback, the avatars were revised after each iteration by adapting, refining, or replacing visual elements to better align with researchers' feedback. For example, feedback included that a magic wand reflects expertise regarding the creation of a new artifact more than a pen does. We further adopted a flat, faceless illustration style inspired by Pablo Stanley's people library [1] to avoid Uncanny Valley effects [28]. We also agreed to color the character gray to prevent stereotypes and to direct participants' attention toward the colorful symbolic elements rather than the character itself. After two co-design iterations, we converged on a final set of 10 aligned avatars and 1 neutral avatar, which are shown in Figure 3. Table 1 provides an overview of the 11 avatars, their visual cues, and the intended expertise in line with common tasks in knowledge work [4].

3.2 Evaluation

To evaluate whether the iteratively designed avatars can be recognized and convey only their intended area of expertise, we conducted an online experiment. To do so, we used a quantitative research approach to statistically test our hypothesis. For this, we collected structured data on participants' assessments of avatars and areas of expertise using LimeSurvey[2].

3.2.1 Study design. We used a within-subjects design in which each participant evaluated all 11 avatars for all 10 areas of expertise in random order, allowing for direct comparisons while controlling for individual biases and minimizing sequence and learning effects. To assess whether an avatar represents a single area of expertise, we state the following null hypothesis: $\mathbf{H_0}$**:** There is no significant difference in participants' expertise ratings for a given avatar. The avatar does not convey ***only one*** specific area of expertise.

3.2.2 Participants. We recruited 100 participants via Prolific[3], applying demographic quotas to ensure a balanced sample (50% male, 50% female) and an age range between 18 and 65 years.[4] We did not apply additional filter criteria, for example regarding MASs knowledge or AI expertise, because we aimed to mirror real organizational settings where knowledge workers exhibit diverse skill levels. Crowd workers inherently perform tasks that involve cognitive effort and computer interaction, similar to knowledge workers in current organizational settings [7]. All participants were residents of European Union countries, a deliberate choice to align the cultural context of the avatar cues with that of the research team [35]. Participants received monetary compensation via Prolific, corresponding to an average reward of £10.86 per hour. Participation was voluntary, and participants could withdraw at any time.

3.2.3 Experimental protocol. After a short welcome message, participants were instructed to evaluate avatars solely on the basis of their appearance. Each participant was sequentially shown all 11 avatars in randomized order. For each avatar, participants rated the extent to which it appeared to embody expertise in each of ten predefined areas of expertise on a 5-point Likert scale (*Strongly disagree–Strongly agree*). The statements are shown in Table 1. In addition, participants could provide open-ended comments, suggest further areas of expertise beyond the predefined list, or critique the avatar's design. To ensure data quality, three attention checks and three comprehension checks were embedded throughout the experiment. The experiment concluded immediately after the evaluation. The total experiment took about 15 minutes per participant.

3.2.4 Data collection & analysis. We collected data from two sources: (1) participants' Likert-scale ratings of each avatar across the ten predefined areas of expertise and (2) embedded attention and comprehension check items to ensure data quality. In total, 100 participants took part in the study, resulting in 100 data sets. Of these, 4 were incomplete, 1 participant failed more than two attention and comprehension checks, and 4 participants were classified, in line with prior HCI research, as speeders (completion time < 50% of the median) [15]. Following these exclusion criteria, 9 data sets were removed, resulting in a final sample of $N = 91$ valid responses. To assess whether each avatar conveys a selective area of expertise, we analyzed differences in perceived expertise across the rated areas (E1–E10) within each avatar. Conceptually, this approach treats expertise as a within-subject factor nested within each avatar and aligns with our goal of testing whether an avatar exhibits a non-uniform (i.e., selective) expertise profile rather than being rated equally across all areas. Given the ordinal nature of the data and the within-subjects design, we employed non-parametric Friedman tests separately for each avatar to examine whether expertise ratings differed across areas. A significant effect indicates that the

[1] https://www.humaaans.com/
[2] https://www.limesurvey.org/
[3] https://www.prolific.com/
[4] The chosen age range covers the EU definition of the core working-age population (20–64 years; Eurostat, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Population_structure_and_ageing).

Figure 3: Ten avatar designs using semiotic cues (e.g., icons and props) to signal agent expertise in common knowledge-worker tasks, following Brachman et al. [4] and one neutral avatar.

avatar is not perceived uniformly across expertise areas, which is a prerequisite for conveying a distinct specialization. To identify whether a specific area of expertise was rated higher than others, we conducted pairwise Wilcoxon signed-rank tests with Holm–Bonferroni corrections, comparing each area against all remaining areas within the same avatar. In addition, we conducted Bayesian repeated-measures ANOVAs for each avatar to quantify the strength of evidence for overall differences across expertise areas (reporting BF_{10}), complemented by pairwise Bayesian t-tests to assess differences between specific areas. For each Friedman test, we report the chi-square statistic χ^2, the corresponding p-value, and Kendall's W as an effect size. To complement these analyses, we calculated Shannon entropy scores to quantify how concentrated or dispersed judgments were across the Likert scale.

All analyses were conducted using JASP[5] (Version 0.95) and the Pingouin[6] package for Python (Version 0.5.5). For transparency and reproducibility, we provide our analysis scripts and the raw data in the supplementary materials.

3.2.5 Results. Figure 4 shows the mean participant ratings for all 11 avatars across all areas of expertise. We conducted Friedman tests to examine differences in participants' expertise ratings across the ten predefined areas for each avatar. As shown in Table 3, all avatars yielded significant results. Bayesian repeated-measures ANOVAs provided converging evidence for these differences, with Bayes factors (BF_{10}) indicating strong to extreme support for the alternative model. Table 2 summarizes the post-hoc analyses of avatars with their top-rated area of expertise. Entries are bolded when an avatar's intended expertise does not match its top-rated expertise. Across Wilcoxon signed-rank tests, avatars A2, A3, A4, A6 and

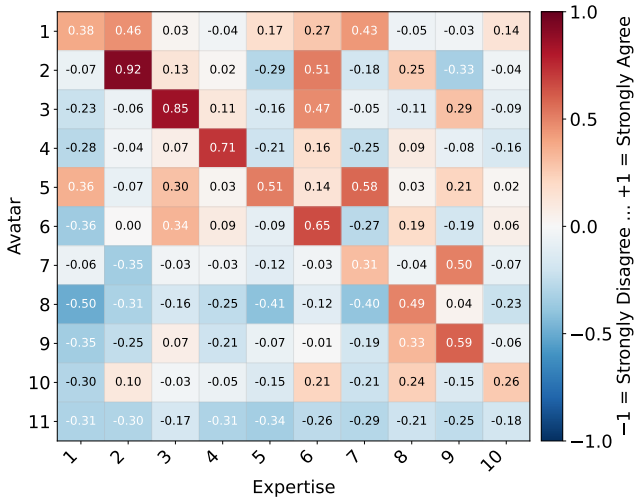

Figure 4: Agreement heatmap of participant mean ratings for avatar–expertise associations.

A8 achieved nine significant comparisons each, with adjusted p-values approaching zero. Bayesian analyses confirmed these effects with Bayes factors (BF_{10}) exceeding 10^4 in all cases, providing very strong evidence. Distributional measures indicated high median ratings of 1.0 for A2, A3, A4, A6, whereas A8 showed a lower mean of 0.5, together with a higher entropy. Avatar A9 produced nine significant Wilcoxon comparisons, slightly lower yet still strong Bayesian support ($BF_{10}^{\text{min}} = 2.5 \times 10^1$), and comparatively higher entropy. The remaining avatars did not show a significantly higher rating in any single area of expertise.

[5]https://jasp-stats.org/
[6]https://pingouin-stats.org/

Table 1: Overview of the 11 avatars, their visual cues, and the intended expertise description based on Brachman et al. [4]

Avatar	Visual Cue / Description	Expertise	Description
Creation			
— *Artifact*			
A1	Holding a wand with sparkles next to a document	E1	Generate a new artifact to be used directly or with some modification
— *Idea*			
A2	Has a thought bubble containing a light bulb	E2	Generate an idea, to be used indirectly
Information			
— *Search*			
A3	Holding a magnifying glass over an open book	E3	Seek a fact or piece of information
— *Learn*			
A4	Pointing at a chalkboard with written items	E4	Learn about a new topic more broadly
— *Summarize*			
A5	Holding multiple highlighted documents in one hand and a shorter document in the other hand	E5	Generate a shorter version of a piece of content that describes the important elements
— *Analyze*			
A6	Pointing at charts and data	E6	Discover a new insight about information or data
Advice			
— *Improve*			
A7	Holding an old document marked with an F and a new document marked with an A	E7	Generate a better version
— *Guidance*			
A8	Holding old scales	E8	Get guidance about how to make a decision
— *Validation*			
A9	Holding a checklist with ticks and crosses	E9	Check whether an artifact satisfies a set of rules or constraints
Automation			
— *Automation*			
A10	Holding a notebook displaying processes	E10	Complete a task in a piece of software with less or no human effort
Neutral			
A11	Neutral avatar without visual cue	E11	No intended expertise

Table 2: Post-hoc results showing each avatar with its top-rated area of expertise. Reported are the number of significant pairwise comparisons (Wilcoxon/Bayesian) for that expertise, as well as distributional measures (mean rating and entropy).

Avatar (Intended Expertise)	Top-rated Expertise (Short Description)	# Sig. Comparisons (Wilcoxon / Bayesian)	Mean	Entropy
A1 (Creation — Artifact)	E1 (Creation — Artifact)	6 / 6	0.38	0.841
A2 (Creation — Idea)	E2 (Creation — Idea)	9 / 9	0.92	0.265
A3 (Information — Search)	E3 (Information — Search)	9 / 9	0.85	0.405
A4 (Information — Learn)	E4 (Information — Learn)	9 / 9	0.71	0.633
A5 (Information — Summarize)	**E7 (Advice — Improve)**	8 / 8	0.58	0.730
A6 (Information — Analyze)	E6 (Information — Analyze)	9 / 9	0.65	0.689
A7 (Advice — Improve)	E7 (Advice — Improve)	8 / 8	0.31	0.877
A8 (Advice — Guidance)	E8 (Advice — Guidance)	9 / 9	0.49	0.815
A9 (Advice — Validation)	E9 (Advice — Validation)	9 / 8	0.59	0.751
A10 (Automation — Automation)	**E6 (Information — Analyze)**	6 / 6	0.21	0.887
A11 (Neutral)	**E3 (Information — Search)**	1 / 1	-0.17	0.808

To evaluate H_0, we combined evidence from global and pairwise tests. Friedman tests indicated significant differences in expertise ratings for 11 of the 11 avatars (Table 3), suggesting that the avatars were not judged uniformly across areas of expertise. Bayesian analyses provided converging evidence for these differences. However, for Avatar 11, the effect size was negligible (Kendall's $W = .026$), suggesting that participants did not consistently attribute a specific area of expertise to this neutral avatar. This supports our design methodology and grounding in semiotic theory, because the expertise cues mainly drove the user's perception of an avatar. Post-hoc Wilcoxon tests with Holm–Bonferroni correction further demonstrated that six avatars displayed a consistent area of expertise, with one expertise significantly outperforming all others (Table 2). This area of expertise aligned with the intended area of expertise for the top six avatars. Bayesian evidence (BF_{10}) provided decisive support for Avatars A2 (Creation — Idea), A3 (Information — Search), A4 (Information — Learn), A6 (Information — Analyze), and A8 (Advice — Guidance), while evidence for A9 (Advice — Validation) was only moderate. In contrast, the remaining avatars did not show a significantly higher rating in any single area of expertise. Accordingly, we reject H_0 for A2, A3, A4, A6, A8, and A9.

Summary. Our online experiment revealed that six out of the ten specialist avatars conveyed a unique area of expertise in knowledge work through the designed expertise cues. The remaining avatars were not rated significantly higher in any single area of expertise, making them less distinguishable and more easily confusable for knowledge workers like Tina.

4 Mixed-Method Study: The Impact of Avatar Design Variants in MAS

Our goal is to enhance explainability in MASs for knowledge workers like Tina while minimizing cognitive load. Having obtained a validated set of avatars from the previous study, we next investigate how they can be embedded as explanations in MASs.

4.1 Designing Explainable MAS with Avatars

Based on the existing literature, we designed and developed four fully functional MAS prototypes with explanation interfaces across two dimensions: explanation modality (text vs. avatar) and explanation resolution (low vs. high). The four prototypes are depicted in Figure 5. To minimize potential confounds, all prototypes contained a working, simple, and scrollable chat interface comparable to currently available chat-based AI interfaces. This design follows established Human-AI interaction guidelines by ensuring that explanations appear in direct conversational context (G4) and remain accessible together with the conversation over time (G12) [2].

High-resolution textual explanations. To help knowledge workers like Tina understand which agents contributed exactly what to a MAS output, one design solution is to provide detailed textual explanations. Similar to how displaying CoT sequences reveal step-by-step reasoning, a detailed step-by-step textual explanation can be used to explain which agents were involved and what their individual contributions were [36]. In the prototype (see Figure 5), these explanations are presented as a detailed textual description displayed in a visually distinct blue text block above the response, explicitly indicating which agents were used and for what purpose.

Low-resolution textual explanations. To assist knowledge workers like Tina in identifying which agents were involved in a MAS output, without being overloaded by detail, textual explanations can be kept intentionally brief. In contrast to the high-resolution variant, which explains both which agents were used and for what purpose, the low-resolution variant specifies only which agents were used [29, 43]. In the prototype (see Figure 5), this is implemented as a shorter version of the same visually distinct blue text block above the system response, indicating only which agents were used.

High-resolution avatar explanations. To enable knowledge workers like Tina to verify which agents contributed to a response without long textual explanations, high-resolution avatar explanations can be used. In a multi-agent chat design, avatars are displayed next to the individual agent outputs (see Figure 5), allowing users to see both the involved agents and their specific contributions [29, 37].

Low-resolution avatar explanations. To allow knowledge workers like Tina to quickly grasp which agents were involved without being overloaded by text or detail, agent participation can also be conveyed non-textual in a low-resolution form. In this variant, avatars are displayed only next to the overall MAS output. Building on prior work that merges textual outputs of individual agents into a single MAS response to reduce cognitive load [29, 43], we extend this idea by visually merging avatar cues (see Figure 5). This variant aims to indicate which agents contributed, without highlighting their individual contributions [37].

In the next step, we wanted to investigate how the different avatar design variants affect interactions with a working MAS. To do so, we conducted a mixed-method study consisting of an experiment and follow-up interviews with a randomly selected subset of participants, assessing how explanation modality (text vs. avatars) and explanation resolution level (low vs. high) affect users' perceived explainability, trust, and cognitive load.

4.2 Study Design

We used a between-subject design in which each participant completed four knowledge working tasks using only one of the four prototypes (four conditions). This design ensured that results were not confounded by comparisons between conditions and prevented learning or carryover effects. To complement the experimental study and gain deeper insights into participants' reasoning, we conducted post-study semi-structured interviews with a randomly selected subset of participants.

4.3 Hypothesis

Prior work shows that explanations support explainability and trust calibration, but can also increase cognitive load when presented in high-resolution form [17, 29, 43]. Visual explanations such as avatars could offer a lightweight alternative that can improve explainability and reduce cognitive load [32, 38]. Building on this, we hypothesize effects of explanation modality (avatars vs. text) and explanation resolution (high vs. low) on three dependent variables: perceived explainability, trust calibration, and cognitive load.

- **H_{1a}:** Participants exposed to avatars will report higher perceived explainability than participants exposed to textual explanations.

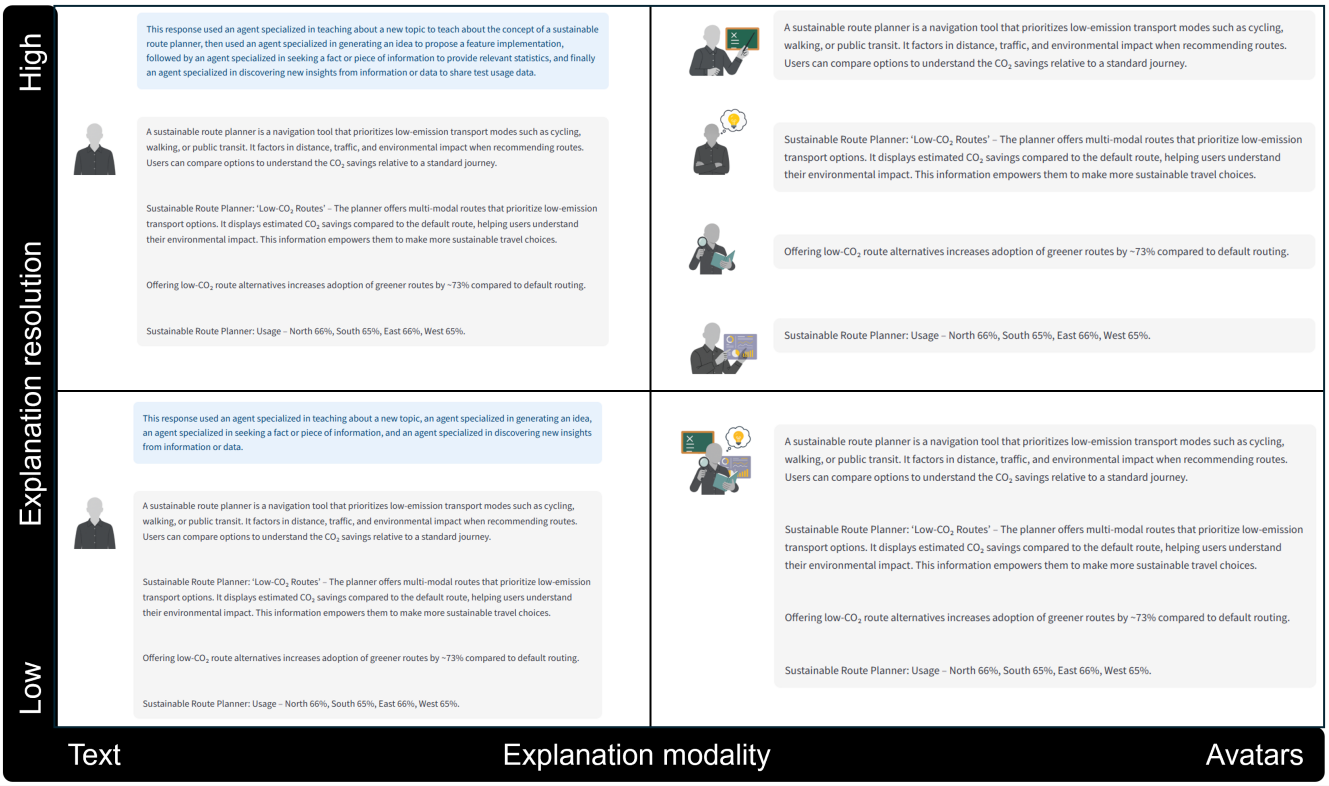

Figure 5: Our four MAS design variations combining explanation modality (text vs. avatars) and explanation resolution (low vs. high).

- **H_{1b}:** Participants exposed to avatars will report higher trust than participants exposed to textual explanations.
- **H_{1c}:** Participants exposed to avatars will report lower cognitive load than participants exposed to textual explanations.
- **H_{2a}:** Across both modalities, participants exposed to high-resolution explanations will report higher perceived explainability than participants exposed to low-resolution explanations.
- **H_{2b}:** Across both modalities, participants exposed to high-resolution explanations will show higher trust than participants exposed to low-resolution explanations.
- **H_{2c}:** Across both modalities, participants exposed to low-resolution explanations will report lower cognitive load than participants exposed to high-resolution explanations.

4.4 Research Approach

We used a mixed-methods approach, combining quantitative and qualitative measures. We collected quantitative data using LimeSurvey, where participants completed standardized questionnaires including the Explanation Satisfaction Scale (ESS) [17], Trust Scale for the XAI Context (TAI) [17] and NASA Task Load Index (NASA-TLX) [14]. In addition, we conducted post-study semi-structured interviews with a randomly selected subset of participants.

4.5 Participants

Similar to the first study, we again chose crowd workers as a proxy for knowledge workers because of the similarities mentioned earlier. We recruited 124 participants via Prolific, applying an age range between 18 and 65 years. Participants from the first experiment were not allowed to take part in the second experiment. Participants were allocated to four different conditions via Prolific's random assignment, which did not allow us to enforce gender quotas. As a result, the final sample (after data cleaning) consisted of 82 males, 38 females, and 1 non-binary participant. Again, all participants were residents of European Union countries to align the cultural context of the avatar cues with that of the research team [35]. Again, we did not apply additional filter criteria, for example regarding MASs knowledge or AI expertise, because we aimed to mirror real organizational settings where knowledge workers exhibit diverse skill levels. After completing the tasks and questionnaires, a random subset of participants took part in a follow-up interview, resulting in 20 semi-structured interviews (5 per condition). Participants received a basic monetary compensation via Prolific, corresponding to an average reward of £9.02 per hour, plus a performance-based reward of up to £2. In addition, participants who took part in a follow-up interview received an extra reward of £3. The study was approved by the University's Ethics and Data Protection Board.

4.6 Apparatus: Multi-Agent System

We developed one fully functional MAS in Python, which was presented to participants in one of four different working Streamlit[7]-based prototypes. As underlying LLM we used the Meta-Llama-3.1 8B Instruct [24]. Based on the results of the first study, we included four agents whose avatars had received the strongest and most unambiguous recognition with one distinct area of expertise (A2, A3, A4, and A6). In addition, we included a fallback agent (A11) that allowed the LLM-based orchestrator to interact with participants outside of specific domains of expertise, for example, when handling social input such as 'thank you' or managing error messages. The limited amount of five agents is in line with previous work that considers five agents practical and manageable in studies investigating MASs [37].

4.7 Tasks

Based on the implemented agents, we designed four tasks grounded in prior research to ensure realistic scenarios and to encourage engagement with multiple agents [4]. The four tasks developed can be found in Table 7. All tasks were set in an open format, ensuring that participants had to explore and interact with the MAS without predefined constraints. Each task could be solved by using a different combination of agents, and participants were incentivized with a £2 bonus to ensure serious interaction with the system in solving the tasks.

4.8 Experimental Protocol

Participants first completed an initial questionnaire to provide demographic information, followed by a short introduction to the tasks and the MAS. Before solving the tasks, participants also completed an initial trust assessment. For this, we used the TAI [17], with minor adaptations to wording to reflect the pre-use context, while maintaining the original meaning of the items. For transperancy, we provide the adapted TAI items in the appendix (see Table 9). Participants were then instructed to only use the working MAS prototype provided by us to solve the tasks. Participants then solved the four tasks in randomized order with one MAS condition through chat. Here, participants were not subject to any time constraints and could decide for themselves when to proceed to the next task. After that, participants rated their trust in the MAS using the unadjusted TAI, and their perceived cognitive load using the NASA-TLX [14]. For transparency, we provide the full items in the appendix (see Table 8 and Table 9). Then participants also completed the ESS [17]. We adapted the wording of the items to align them with our study context while preserving their original meaning. For transparency, we provide the adapted ESS items in the appendix (see Table 10). Finally, we conducted semi-structured interviews with a random subset of participants to ensure that the qualitative findings were not limited to a specific subgroup, thereby supporting external validity. The interview guide is provided in the appendix (see Table 11). The median time for solving the four tasks and completing the questionnaires was about 30 minutes, and additional interviews lasted around 20 minutes.

[7] https://www.streamlit.io/

4.9 Data Collection & Analysis

We collected data from four sources: (1) participants' questionnaire responses in LimeSurvey, including ESS, TAI, and NASA-TLX, (2) three embedded attention and comprehension check items to ensure data quality, (3) qualitative data from the semi-structured interviews ($N = 20$), (4) and prototype interaction data. In total, we obtained 124 data sets. Of these, two participants failed two or more attention checks, and one entry was excluded due to missing Prolific identification. Following these exclusion criteria, the final sample comprised $N = 121$ valid responses. Participants were distributed approximately evenly across the four conditions (low-resolution textual: 31, low-resolution avatar: 28, high-resolution textual: 32, high-resolution avatar: 30). For the analysis, we focused on changes in trust by calculating deltas between the pre- and post-interaction TAI scores. We analyzed these together with the responses of ESS and NASA-TLX by using Kruskal–Wallis tests. To complement the non-parametric analyses, we report Bayesian factors (BF_{10}). All analyses were conducted using JASP (Version 0.95) with default priors. The first author conducted a reflexive thematic analysis following Braun and Clarke's six-phase approach [5]. The first author conducted the qualitative analysis, beginning with familiarization and open, inductive coding. Themes were iteratively developed and refined through repeated engagement with the data. Following Braun and Clarke [6], reflexive thematic analysis does not require multiple coders; hence, the process was carried out by the first author. To ensure rigor, codes and themes were revisited after phases of distancing and engagement with the dataset.

4.10 Results

To structure our results, we first report on the interaction times and then focus on the effects of explainability. Specifically, we examine participants' satisfaction with the explanations and how these shaped the formation of their mental models regarding how the MAS generates results, in particular their understanding of which agents were involved and what contributions they made [16, 23]. We then examine the effects on trust, emphasizing how participants calibrated their trust according to their condition. Finally, we analyze the perceived cognitive load across the conditions.

At the quantitative level, we did not observe significant differences between conditions across explainability (ESS), trust (TAI), or cognitive load (NASA-TLX) (all $p > .10$, $\varepsilon^2 < .05$, see Table 4, Table 5, and Table 6). Bayesian analyses consistently favored the null model ($BF_{10} < 0.50$), suggesting that there were no significant differences across trust, but also confirmed that cognitive load remained comparable between conditions. Reliability analyses confirmed acceptable to good internal consistency for adjusted TAI and ESS scales (trust-deltas: $\alpha = .774$, 95% CI [.719, .830]; ESS: $\alpha = .889$, 95% CI [.853, .925]), supporting the robustness of the measures. Detailed reliability results for all scales are provided in the supplementary material. Given the lack of quantitative effects, we report pooled descriptive statistics and qualitative findings, which provide first insights into how participants perceived the different explanation designs.

4.10.1 MAS Interaction Time. To verify that participants chatted diligently and conformed to their assigned condition, we analyzed task interaction times. After outlier removal using the box-plot

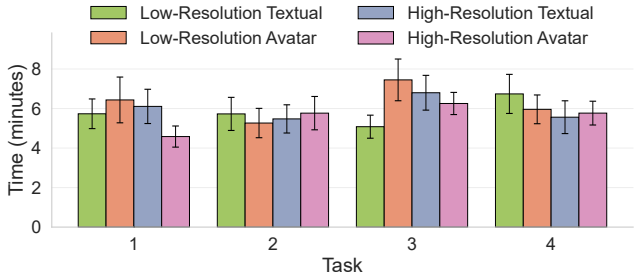

Figure 6: Mean task interaction times (in minutes) across the four conditions, shown separately for Tasks 1–4. Error bars indicate ±SE after outlier removal.

method, average task interaction times (defined as the duration from the first user message to the last MAS message within a task) ranged between 4.6 and 7.5 minutes (see Figure 6). For Task 1, mean interaction times varied between 4.6 min (high-resolution avatar) and 6.4 min (low-resolution avatar), while Task 2 and Task 3 showed slightly longer interaction times in some conditions (e.g., 7.5 min in low-resolution avatar). Task 4 again resulted in comparable interaction times across conditions (approximately 5.6 – 6.7 min). Standard errors were small, and error bars overlapped in all cases, indicating no substantial differences in task interaction times across conditions. This confirms that participants engaged with the working MAS diligently and conformed across conditions.

4.10.2 Effects on Explainability. We report pooled descriptive statistics across conditions (1 = strongly agree, 5 = strongly disagree). Ratings indicated more neutral judgments that all conditions aided understanding of how responses were generated ($M = 2.60$, $SD = 0.95$; $Md = 3$), and agreement that they were satisfying ($M = 2.50$, $SD = 0.97$; $Md = 2$). Participants judged that explanations provided sufficient detail ($M = 2.55$, $SD = 0.99$; $Md = 2$) and were reasonably complete ($M = 2.59$, $SD = 1.10$; $Md = 2$). They also indicated that explanations supported effective system use ($M = 2.50$, $SD = 0.98$; $Md = 2$) and were useful for achieving task goals ($M = 2.33$, $SD = 0.93$; $Md = 2$). Participants were more neutral about using the explanations to judge the system's accuracy ($M = 2.76$, $SD = 1.03$; $Md = 3$). However, our thematic analysis provided more insights into how the different conditions shaped participants' experiences.

Explanation satisfaction. Participants expressed mixed reactions to the different explanation formats. Low-resolution textual explanations were sometimes appreciated for their clarity and minimalism: "I like the minimalism of this and I don't want to be overwhelmed with information" (P3) and other times dismissed as superficial or ignored: I saw what it said that it uses a special agent for this, but not really paid much attention to it" (P2). Similarly, merged avatars in the low-resolution avatar condition were noticed but often regarded as decorative and of limited informational value: "I didn't use that difference on the drawings or whatever to help me. So I just was neutral about that" (P6). High-resolution textual explanations provided more transparency for some participants: "So then it's easier for me to find that specific information probably" (P13) while others "blended it out" (P15). By contrast, separate avatars next to each contribution were widely perceived as useful, allowing participants to quickly identify the type of information provided: "It could guide me, for example, [...] I just needed some fresh ideas on it, then I would [...] go directly to the second answer based on the avatar [...] that would help me save time" (P16).

Mental models of agent involvement. Across conditions, participants differed in their understanding of the MAS. In the low-resolution textual condition, most interpreted the system as a single agent: "I felt like I was talking to the same person"(P3). Similar impressions emerged with merged avatars, where the symbolic cues (e.g., light bulb for idea generation) were noted but rarely led to a strong sense of multiple agents: "When I saw the light bulb, I thought that was giving new ideas. That's what I interpreted" (P6) and "It sounded the same like one person" (P8). High-resolution textual explanations increased awareness of multiple agents: "I think there were multiple because it says like [...] this response used an agent specialized [...]" (P13) while others still perceived the system as a single agent: "It felt like one. Because the tone was about the same like speech wise just kind of the same [...]" (P11). The clearest differentiation arose in the high-resolution avatar condition where avatars helped participants describe outputs as distinct perspectives: "I got like different perspectives and then I could choose a little bit if I wanted the data analytics perspective for the question or just a general [...] summary of the topic" (P20). Still, even here some participants framed the system as a single entity with "different personalities" (P17).

4.10.3 Effects on Trust. We report pooled descriptive statistics for the trust deltas (−4 to +4 per item; note that negative values indicate higher agreement). Overall changes in trust were minimal ($M = 0.11, SD = 0.21; Md = 0.00$). Items with slightly more positive shifts included perceived efficiency ($M = 0.18, SD = 1.15$) and feeling safe in relying on the system ($M = 0.21, SD = 1.02$), while agreement with being wary of the system also increased slightly ($M = -0.22, SD = 1.22$).

Trust calibration. Nevertheless, qualitative data suggested that explanations across conditions shaped how participants calibrated their trust in the MAS. In the low-resolution textual condition, some participants described the short explanation as a reminder to remain cautious, noting that it helped "keeping in mind that the answers might not be completely accurate [...]" (P1). Others reported greater confidence when multiple agents were mentioned: "I felt more confident in trusting the answer when there was several agents" (P3). By contrast, merged avatars in the low-resolution avatar condition were mostly perceived as neutral cues, with participants emphasizing that they did not affect trust because they were seen as decorative: "They did not decrease or increase my trust [...] because I was not even paying a lot of attention to them" (P9). High-resolution textual explanations produced mixed effects. For some, knowing that a specialized agent contributed provided confirmation: "it gives you like a sort of a confirmation that you can trust the information when you know it's a specialized agent" (P13). Others stated that it did not influence trust because they "blended it out" (P15). Distinct avatars in the high-resolution avatar condition had the clearest influence on trust, as several participants described

greater confidence when multiple perspectives were visible: "the different pictures helped to trust the system in general because I was like if all the angles check out then [...] it's probably true" (P20). At the same time, other participants emphasized that avatars were simply an additional cue and did not meaningfully affect trust: "I mean, the avatars are just an extra cue, nothing else to me" (P16).

Agent-dependent trust. In addition to calibration effects, some participants expressed trust that varied depending on the specific agent involved. For example, in the low-resolution textual condition, one participant reported greater confidence in MAS outputs that used a factual agent: "When it said that it's using like factual information agent, I more trusted the agent [...]" (P4). In the high-resolution text condition, the explicit mention of specialization reinforced trust: "it gives you like a sort of a confirmation that you can trust the information when you know it's a specialized agent" (P13). Similarly, in the high-resolution avatar condition, one participant emphasized reliance on particular agents: "Then, of course, I knew in the third one that I could rely on it as well" (P20).

4.10.4 Effects on Cognitive Load. First, we report descriptive statistics for the raw (unweighted) NASA-TLX (0–20 per subscale; lower = lower workload; Performance reverse-scored so higher = greater cost). Overall workload (mean of subscales) was moderate ($M = 9.49, SD = 3.62; Md = 9.67$). The largest contributors were Mental Demand ($M = 12.43, SD = 5.24; Md = 14$) and Effort ($M = 11.99, SD = 5.06; Md = 12$), followed by Temporal Demand ($M = 10.78, SD = 5.87; Md = 12$) and perceived Performance cost ($M = 9.57, SD = 4.76; Md = 10$). Frustration was lower ($M = 8.48, SD = 5.86; Md = 8$), and Physical Demand was minimal across conditions ($M = 3.72, SD = 5.24; Md = 1$).

Minimalism & ignorability. Low-resolution explanations, whether textual or avatars, were typically seen as unobtrusive and easy to ignore. Participants noted that they focused primarily on the MAS output, describing the minimalism as preferable: "I like the minimalism of this and I don't want to be overwhelmed with information" (P3), while others emphasized that "what matters is the content of the message. So I don't think it makes that difference" (P1). Similarly, in the merged avatar condition, participants dismissed the visual cues as neutral: "no, they did not bother me" (P9) or admitted to "not paying much attention to the pictures" (P10).

Confusion vs. support. In contrast, high-resolution textual explanations were more likely to introduce confusion. Several participants described being unsure whether they were reading an explanation or a response: "I found myself confused, like, okay, is this my response or is this [...] just the information about the agent" (P11). Another participant criticized the detailed explanation text as distracting: "too much text in addition to the messages" (P14). Here, participants suggested to make "it stand out more" (P11) and to "make it more visual [...] like arrows to show the interaction" (P15). Other participants admitted that they simply ignored it: "I mean, in the task, I didn't pay attention to it" (P12). Still, a minority of participants found the explanation "clear to read" (P13). In contrast, the distinct avatars in the high-resolution avatar condition were generally experienced as supportive rather than cognitively demanding. Participants emphasized that the avatars helped them focus and distinguish between contributions: "I think having just the text and the images, it's pretty neat. It kind of helps you keep focused. Not too much is happening on the screen" (P16). Another participant highlighted the ease of use: "I found the system very easy, useful, simple to understand. I liked the pictures. It gave clear and quick answers about what I was asking for" (P18). Another participant stated that its initial confusion was quickly resolved: "In the beginning, I was totally confused that I got not one response, but four or five. But that was quickly resolved" (P20).

4.10.5 Summary. Consistent interaction times across tasks and conditions suggest that participants engaged diligently and conformed to the assigned condition throughout the study. Across all conditions, participants' perceptions of the different explanation designs appeared to shape their perceived explainability, trust, and cognitive load in different ways. Quantitative results did not reveal significant differences between conditions; therefore, we cannot clearly accept or reject our hypotheses. However, qualitative findings provide initial indications of differences, and we address potential reasons for this discrepancy in the discussion section (5.3). As a reminder, H1 hypotheses address the effects of modality (avatars vs. text), H2 hypotheses address the effects of explanation resolution (high vs. low), and a, b, and c hypotheses refer to perceived explainability, trust, and cognitive load, respectively. For **H**1*a*, we found tentative qualitative support. While low-resolution merged avatars were often described as decorative and did not substantially improve participants' understanding of agent involvement, high-resolution distinct avatars were perceived as helping participants distinguish between different agents and contributions, which may have supported more accurate mental models than participants in textual explanation conditions. For **H**1*b*, the evidence is qualitatively inconclusive. Low-resolution avatars had little to no influence on trust. High-resolution avatars, however, were sometimes described as leading to higher trust compared to textual explanations. For **H**1*c*, we found preliminary qualitative support. Both low-resolution and high-resolution avatars were typically described as unobtrusive, whereas high-resolution textual explanations sometimes caused confusion or distraction. For **H**2*a* we found tentative qualitative support. High-resolution avatars seemed to support participants' understanding of agent specialization, whereas high-resolution textual explanations were more ambivalent: some participants benefited from the transparency, while others ignored or found them confusing. For **H**2*b* we also found tentative qualitative support. Trust appeared to increase when specialized agents were made visible in the high-resolution avatar condition, but high-resolution textual explanations were perceived as helpful by some participants and distracting by others. Finally, **H**2*c* is cautiously qualitatively supported. Low-resolution explanations, whether textual or avatars, were easily ignored and added little cognitive load, as supported by our quantitative and qualitative analysis. High-resolution textual explanations, in contrast, occasionally introduced confusion and distraction, while high-resolution avatars were perceived as informative while adding little to cognitive load.

5 Discussion

In the following, we discuss our findings and limitations in two parts: first, we reflect on our avatar designs, which investigated how users perceive avatar cues in relation to areas of expertise. Second,

we turn to our MAS experiment, which examined how avatars and textual explanations shape explainability, trust, and cognitive load.

5.1 Designing Avatars with Expertise Cues

Visual cues can reveal agent expertise. Our results show that avatars A2, A3, A4, A6, A8, and A9 received significantly higher ratings in their designed area of expertise. Although our analysis did not require that avatars be rated highest in their intended area of expertise, participants nevertheless perceived them as clearly representing their intended area of expertise, which corroborates the effectiveness of our design. These results suggest that visual cues in avatars can effectively communicate their agents area of expertise, even when participants are not instructed to make such associations. Knowledge workers like Tina can therefore identify an agent's expertise solely from its avatar, provided that the avatar is designed to represent the agent's specialization. Prior research has centered on avatar design for general-purpose or social interaction contexts [32]. Our results extend this work by showing that avatars can signal areas of expertise based on Brachman et al. [4]'s LLM-use subcategories in knowledge work. In doing so, our work provides empirical evidence for selected avatars and outlines a methodological approach for evaluating avatar–expertise alignment for specialized agents. Nevertheless, avatars A1, A5, A6, and A10 did not receive significantly higher ratings in only one area of expertise compared to all others. A possible explanation is that these avatars appeared more generic, for example, by simply holding sheets of paper or a laptop, which could suggest a range of different areas of expertise. To improve agent expertise recognition for knowledge workers, we argue that avatar design should therefore employ domain-specific and distinctive visual cues, for example, tools, attire, or symbolic elements that are strongly associated with a given area of expertise. From a semiotic perspective, recognition improves when signifiers map transparently and unambiguously to the intended signified [30]. Designers should also ensure that these cues are mutually distinctive across avatars, so that knowledge workers can easily differentiate between areas of expertise without ambiguity. Future work could build on these less distinctive designs to develop and evaluate avatars that address the remaining subcategories (E1, E5, E6, and E10) of knowledge-workers' use of LLMs as identified by Brachman et al. [4].

Beyond static images. Our avatar character and the visual cues were static and generated with a generative AI tool, which resulted in occasional inconsistencies and imperfections. We did not examine avatar customization in this study, as it would have introduced confounding variables and may have shifted the focus from testing recognition of expertise to, e.g., questions of ownership. Future work could enhance these avatars with professionally designed characters and visual cues that ensure visual coherence and explore how customization (e.g., character, colors or cues) impacts agent expertise recognition for knowledge workers like Tina. Prior work by Ha et al. [13] suggests that customized agent personas strengthen emotional bonds and trust compared to generic agents. Yet avatar personalization risk reinforcing stereotypes, as highlighted by Ratan and Sah [33]. In this study, we also left open how more dynamic representations (e.g., 3D avatars, videos, animations of interactions) may influence agent expertise recognition for knowledge workers. Dynamic representations could be used to convey areas of expertise that are difficult to express through static visuals, for example, solving a task in a piece of software with minimal human effort (E10). Prior work on explainable embodied agents suggests that movement-based cues, such as gestures or gaze can significantly enhance the transparency and intelligibility of an agent's internal state and intentions [42]. Future work could therefore examine how to leverage dynamic avatars without introducing distraction or cognitive overload.

5.2 Avatars for Explainability in MAS

While our quantitative analyses did not reveal statistically significant differences between conditions (see Section 4.10), the qualitative findings help contextualize these null effects and provide exploratory insights into how participants perceived different explanation designs.

***Explainability**: Distinct high-resolution avatars perceived as supporting clearer mental models.* Our qualitative findings provide partial support for H_{1a} and H_{2a}. While low-resolution avatars were often perceived as merely decorative and did not substantially aid participants' understanding of agent involvement, high-resolution avatars were perceived as helping participants to distinguish between different agents and their respective contributions. These were described as supporting the formation of clearer and more accurate mental models compared to textual explanations. However, these perceptions should be interpreted with caution, as they were not reflected in statistically significant differences in the quantitative measures of explainability. This implies *Design Implication (DI) I: MAS designers may consider employing high-resolution avatars to let knowledge workers like Tina differentiate agents and outputs*, as qualitative evidence suggests that they can support clearer mental models than textual explanations. This observation is consistent with and extends prior work on explanation modalities: Szymanski et al. [38] showed that users often preferred visual over textual explanations, even though they sometimes misinterpreted them, while Schelhorn et al. [36] demonstrated that textual CoT displays can increase representational fidelity in LLM-based analytics agents. Building on this line of work, our qualitative results suggest that high-resolution avatars serve as a complementary visual modality that may help users form accurate mental models of MASs.

***Trust**: High-resolution explanations shape confidence in MAS output.* The qualitative evidence for H_{1b} and H_{2b} is mixed, but points toward a potential role of resolution for calibrated trust. Low-resolution avatars had little to no influence on participants' trust, as they were often ignored or dismissed as decorative. This in line with prior work that even suggests that overly simplified explanations can undermine trust [22]. Importantly, these qualitative patterns did not translate into statistically significant differences in trust across conditions, suggesting that these effects may be subtle or context-dependent. This implies *DI II: MAS designers may avoid low resolution explanations*, as qualitative findings suggest they may have limited or even negative effects on knowledge workers' trust. In contrast, high-resolution avatars were often described as increasing confidence, particularly when participants perceived alignment across multiple agents, suggesting that visible consensus may act as

a trust cue. This observation aligns with Song et al. [37], who found that participants were more likely to shift their opinions when multiple agents shared the same stance rather than a single agent. This implies *DI III: MAS designers may consider employing high-resolution avatars to potentially increase knowledge workers' trust*, while remaining aware of potential confounding trust accelerators driven by social influence. High-resolution textual explanations, however, produced ambivalent reactions: while some participants valued the detailed transparency, others described them as distracting or confusing. This implies *DI III: MAS designers may consider using detailed textual explanations selectively based on knowledge workers and context*, as they may either strengthen or undermine trust.

***Cognitive Load**: Avatars remain lightweight, textual explanations pose a risk of overload.* Our qualitative results preliminary support H_{1c} and H_{2c}. Across conditions, participants consistently described both low- and high-resolution avatars as unobtrusive, suggesting that avatars may convey explanations without substantially increasing perceived cognitive load. This aligns with the quantitative results, which showed no significant differences in cognitive load between conditions. In contrast, high-resolution textual explanations occasionally introduced confusion or distraction, suggesting that textual detail may risk overloading users when not carefully structured. This implies *DI IV: MAS designers may consider employing non-textual visual explanations, such as avatars*, to to reduce the risk of cognitive overload of knowledge workers like Tina. Low-resolution explanations, regardless of modality, were often ignored and therefore seemed to add little to cognitive load, but also contributed minimally to participants' understanding (DI II). These findings are consistent with prior work showing that explanation modality influences perceived cognitive effort [22]. Wang et al. [43] demonstrated that detailed textual reasoning in LLM interfaces can overwhelm users, particularly in low-risk contexts, while Szymanski et al. [38] reported that non-expert users often prefer visual explanations despite interpreting them less accurately than text. In contrast, Naik et al. [29] noted that early adopters of MASs were concerned that high-resolution explanations might overwhelm users. [9] argue that transparency should be designed "as good as possible" rather than aiming for exhaustive detail, emphasizing context-specific solutions over universal guidelines. Similarly, our qualitative results preliminary suggest that lightweight visual explanations, such as avatars, may support knowledge workers mental models of MASs without overwhelming them, whereas textual explanations require careful structuring to avoid confusion.

5.3 Limitations and Future Work

Like any other study, our work comes with limitations that should be addressed in future work. Regarding our avatar design research, we focused on static avatars, leaving open how dynamic avatars or customized avatars might affect expertise recognition. Future work should address these limitations by extending avatar designs beyond static representations. Second, our mapping of areas of expertise was grounded in Brachman et al. [4], a categorization not intended as a foundation for distinctive avatar design, which may have constrained category discriminability and the distinctiveness of avatar cues.

Absence of Significant Differences. The first limitation of the MAS experiment is that we were unable to capture significant quantitative results. Our analysis of interaction times indicates that participants engaged seriously with the MAS. Nevertheless, we may not have fully reproduced the real-life pressures of knowledge workers like Tina, where professional performance has tangible consequences. As a result, participants may have engaged less strongly with the provided explanations than knowledge workers whose jobs depend on high-quality decisions. We believe that such motivational dynamics are difficult to capture in an artificial setting, and we recommend that future research conduct longitudinal field studies in authentic work contexts to examine how explanation modalities and resolutions affect knowledge workers. Legal compliance and data protection obligations prevented us from conducting such studies. Second, as we made wording adaptations to the TAI and ESS items to match the pre-use and study context, potential influences on the scales' validity cannot be fully excluded. However, these adaptations were purely linguistic and did not alter the underlying constructs. We verified the internal consistency of these scales using Cronbach's alpha. Reliability was acceptable to good across all measures ($\alpha > .77$), indicating that the scales performed robustly despite linguistic adaptations. Exact reliability coefficients for each scale are reported in the supplementary material. Future work could further examine the robustness of these measures across different contexts. It is also possible that the selected quantitative questionnaires did not fully capture the nuanced perceptual differences observed in the qualitative data, suggesting the value of complementary mixed-methods approaches in future research. Another limitation concerns our reliance on crowd workers. Although this allowed us to recruit a diverse online sample, crowd workers may differ from professional knowledge workers in terms of motivation, domain expertise, and accountability. These differences limit the generalization of our findings, highlighting the need for future studies to investigate explanation modalities and resolutions with professionals.

Scalability. While our research suggests that avatars can support the perception of agent expertise and contribution in small-sized MAS, designing and validating distinctive avatars for a larger number of agents is challenging. As the number of agents grows, visual distinctiveness becomes harder to maintain, and the evaluation effort increases substantially. Future work should therefore explore scalable strategies for representing areas of expertise, for instance, through systematic design frameworks or adaptive visualizations, to ensure that avatar-based explainability remains feasible in larger MAS. Last, our samples consisted of online participants from the European Union, limiting generalization to other cultural contexts and professional knowledge workers. Prior CHI work shows that the interpretation of images, symbols, and colors varies across cultures [35]. As a result, future work should also examine the role of cultural background and domain expertise by conducting cross-cultural and field studies.

6 Conclusion

In this work, we set out to enhance explainability in MASs, enabling knowledge workers to understand which agents contributed to a result. For this 12 researchers co-designed in 2 iterations 11 avatars representing areas of expertise. We showed that avatars can be designed to convey a single area of expertise. This provided a visual basis for representing agent roles without increasing cognitive load. While low-resolution textual explanations or merged avatars were often ignored, qualitative findings suggested that distinct high-resolution avatars next to each agent's output were perceived as supporting explainability and calibrated trust while not adding to perceived cognitive load. At the same time, our findings highlight important trade-offs: textual explanations can risk overwhelming users if not carefully structured, and avatar-based consensus cues may shape trust beyond transparency. We provide validated avatar designs from our first study, qualitative empirical insights on explanation designs from our second study, and concrete design implications for balancing explainability, trust, and cognitive load in MASs. Future research should investigate how avatar-based explanations can scale to larger MAS, adapt to cultural contexts, and move beyond static representations toward more dynamic visualizations that support knowledge workers in practice.

Acknowledgments

This work was supported by Energie Baden-Württemberg AG (EnBW). ChatGPT was used to assist with language improvement.

References

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–18. doi:10.1145/3173574.3174156

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3290605.3300233

[3] Teodoro Baldazzi, Luigi Bellomarini, Stefano Ceri, Andrea Colombo, Andrea Gentili, and Emanuel Sallinger. 2023. Fine-Tuning Large Enterprise Language Models via Ontological Reasoning. In *Rules and Reasoning*, Anna Fensel, Ana Ozaki, Dumitru Roman, and Ahmet Soylu (Eds.). Springer Nature Switzerland, Cham, 86–94.

[4] Michelle Brachman, Amina El-Ashry, Casey Dugan, and Werner Geyer. 2024. How Knowledge Workers Use and Want to Use LLMs in an Enterprise Context. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 189, 8 pages. doi:10.1145/3613905.3650841

[5] Virginia Braun and Victoria Clarke. 2012. Thematic analysis. In *APA handbook of research methods in psychology, Vol. 2. Research designs: Quantitative, qualitative, neuropsychological, and biological*, Harris Cooper, Paul M. Camic, Deborah L. Long, A. T. Panter, David Rindskopf, and Kenneth J. Sher (Eds.). American Psychological Association, 57–71. doi:10.1037/13620-004

[6] Virginia Braun and Victoria Clarke. 2019. Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health* 11, 4 (2019), 589–597. arXiv:https://doi.org/10.1080/2159676X.2019.1628806 doi:10.1080/2159676X.2019.1628806

[7] Thomas Davenport. 2005. Thinking for A Living: How to Get Better Performance and Results from Knowledge Workers. (01 2005).

[8] Ferdinand de Saussure. 1916. *Cours de linguistique générale*. Payot, Paris. Edited by Charles Bally and Albert Sechehaye.

[9] Malin Eiband, Hanna Schneider, Mark Bilandzic, Julian Fazekas-Con, Mareike Haug, and Heinrich Hussmann. 2018. Bringing Transparency Design into Practice. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (Tokyo, Japan) *(IUI '18)*. Association for Computing Machinery, New York, NY, USA, 211–223. doi:10.1145/3172944.3172961

[10] Will Epperson, Gagan Bansal, Victor C Dibia, Adam Fourney, Jack Gerrits, Erkang (Eric) Zhu, and Saleema Amershi. 2025. Interactive Debugging and Steering of Multi-Agent AI Systems. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 156, 15 pages. doi:10.1145/3706598.3713581

[11] Asbjørn Følstad, Cecilie Bertinussen Nordheim, and Cato Alexander Bjørkli. 2018. What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study. In *Internet Science*, Svetlana S. Bodrunova (Ed.). Springer International Publishing, Cham, 194–208.

[12] Anne Fota, Katja Wagner, Tobias Röding, and Hanna Schramm-Klein. 2022. "Help! I Have a Problem" – Differences between a Humanlike and Robot-like Chatbot Avatar in Complaint Management. doi:10.24251/HICSS.2022.522

[13] Juhye Ha, Hyeon Jeon, Daeun Han, Jinwook Seo, and Changhoon Oh. 2024. CloChat: Understanding How People Customize, Interact, and Experience Personas in Large Language Models. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 305, 24 pages. doi:10.1145/3613904.3642472

[14] Sandra G. Hart and Lowell E. Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*, Vol. 52. North-Holland, 139–183.

[15] Franziska Herbert, Steffen Becker, Leonie Schaewitz, Jonas Hielscher, Marvin Kowalewski, Angela Sasse, Yasemin Acar, and Markus Dürmuth. 2023. A World Full of Privacy and Security (Mis)conceptions? Findings of a Representative Survey in 12 Countries. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 582, 23 pages. doi:10.1145/3544548.3581410

[16] Alexandre Heuillet, Fabien Couthouis, and Natalia Díaz-Rodríguez. 2022. Collective eXplainable AI: Explaining Cooperative Strategies and Agent Contribution in Multiagent Reinforcement Learning With Shapley Values. *IEEE Computational Intelligence Magazine* 17 (02 2022), 59 – 71. doi:10.1109/MCI.2021.3129959

[17] Robert R. Hoffman, Shane T. Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* Volume 5 - 2023 (2023). doi:10.3389/fcomp.2023.1096257

[18] Donghan Hu and Sang Won Lee. 2020. ScreenTrack: Using a Visual History of a Computer Screen to Retrieve Documents and Web Pages. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376753

[19] Liss Jenneboer, Carolina Herrando, and Efthymios Constantinides. 2022. The Impact of Chatbots on Customer Loyalty: A Systematic Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research* 17 (01 2022), 212–229. doi:10.3390/jtaer17010011

[20] Zhiqiu Jiang, Mashrur Rashik, Kunjal Panchal, Mahmood Jasim, Ali Sarvghad, Pari Riahi, Erica DeWitt, Fey Thurber, and Narges Mahyar. 2023. CommunityBots: Creating and Evaluating A Multi-Agent Chatbot Platform for Public Input Elicitation. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 36 (April 2023), 32 pages. doi:10.1145/3579469

[21] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 250, 17 pages. doi:10.1145/3544548.3581001

[22] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. Too much, too little, or just right? Ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*. 3–10. doi:10.1109/VLHCC.2013.6645235

[23] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[24] Meta AI. 2024. Meta Llama 3.1 8B Instruct — Model Card. https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct. Model card accessed via Hugging Face, released July 23, 2024.

[25] Microsoft. 2024. AI at Work Is Here. Now Comes the Hard Part. https://www.microsoft.com/en-us/worklab/work-trend-index/ai-at-work-is-here-now-comes-the-hard-part. Microsoft Work Trend Index.

[26] Microsoft. 2024. How the Microsoft 365 Copilot Orchestrator Selects Plug-Ins. https://learn.microsoft.com/de-de/microsoft-365-copilot/extensibility/ Accessed: January 17, 2025.

[27] Takato Mizuho, Takuji Narumi, and Hideaki Kuzuoka. 2024. Investigating the Effects of Changing the Appearance of Screen-Based Avatars on Audience Memory. In *ACM Symposium on Applied Perception 2024* (Dublin, Ireland) *(SAP '24)*. Association for Computing Machinery, New York, NY, USA, Article 7, 9 pages. doi:10.1145/3675231.3675239

[28] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine* 19, 2 (2012), 98–100. doi:10.1109/MRA.2012.2192811

[29] Suchismita Naik, Austin L. Toombs, Ph.D. Snellinger, Amanda, Scott Saponas, and Amanda K Hall. 2025. Designing with Multi-Agent Generative AI: Insights from Industry Early Adopters. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference (DIS '25)*. Association for Computing Machinery, New York, NY, USA, 1961–1972. doi:10.1145/3715336.3735823

[30] Charles Sanders Peirce. 1931. *Collected Papers of Charles Sanders Peirce.* Harvard University Press, Cambridge, MA. Volumes I–VI edited by Hartshorne and Weiss (1931–1935), Volumes VII–VIII by Burks (1958).

[31] Kexin Quan, Dina Albassam, Mengke Wu, Zijian Ding, and Jessie Chin. 2025. Towards AI as Colleagues: Multi-Agent System Improves Structured Professional Ideation. arXiv:2510.23904 [cs.HC] https://arxiv.org/abs/2510.23904

[32] Mashrur Rashik, Mahmood Jasim, Kostiantyn Kucher, Ali Sarvghad, and Narges Mahyar. 2024. Beyond Text and Speech in Conversational Agents: Mapping the Design Space of Avatars. In *Proceedings of the 2024 ACM Designing Interactive Systems Conference* (Copenhagen, Denmark) *(DIS '24)*. Association for Computing Machinery, New York, NY, USA, 1875–1894. doi:10.1145/3643834.3661563

[33] Rabindra Ratan and Young June Sah. 2015. Leveling up on stereotype threat: The role of avatar customization and avatar embodiment. *Computers in Human Behavior* 50 (2015), 367–374. doi:10.1016/j.chb.2015.04.010

[34] Adam Rule, Aurélien Tabard, and Jim Hollan. 2017. Using Visual Histories to Reconstruct the Mental Context of Suspended Activities. *Human–Computer Interaction* 32, 5-6 (2017), 511–558. arXiv:https://doi.org/10.1080/07370024.2017.1300063 doi:10.1080/07370024.2017.1300063

[35] Patricia Russo and Stephen Boor. 1993. How fluent is your interface? designing for international users. In *Proceedings of the INTERACT '93 and CHI '93 Conference on Human Factors in Computing Systems* (Amsterdam, The Netherlands) *(CHI '93)*. Association for Computing Machinery, New York, NY, USA, 342–347. doi:10.1145/169059.169274

[36] Till Schelhorn, Ulrich Gnewuch, and Alexander Maedche. 2025. The Impact of Chain-of-Thought Display on the Effective Use of LLM-based Analytics Agents. In *SIGHCI 2024 Proceedings*. https://aisel.aisnet.org/sighci2024/15

[37] Tianqi Song, Yugin Tan, Zicheng Zhu, Yibin Feng, and Yi-Chieh Lee. 2025. Greater than the Sum of its Parts: Exploring Social Influence of Multi-Agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 309, 11 pages. doi:10.1145/3706599.3719973

[38] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 109–119. doi:10.1145/3397481.3450662

[39] Chek Tien Tan, Indriyati Atmosukarto, Budianto Tandianus, Songjia Shen, and Steven Wong. 2025. Exploring the Impact of Avatar Representations in AI Chatbot Tutors on Learning Experiences. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1072, 12 pages. doi:10.1145/3706598.3713456

[40] Sara Tandon and Jennifer Wang. 2023. Surfacing AI Explainability in Enterprise Product Visual Design to Address User Tech Proficiency Differences. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI EA '23)*. Association for Computing Machinery, New York, NY, USA, Article 398, 8 pages. doi:10.1145/3544549.3573867

[41] Astrid M. von der Pütten, Nicole C. Krämer, Jonathan Gratch, and Sin-Hwa Kang. 2010. "It doesn't matter what you are!" Explaining social effects of agents and avatars. *Comput. Hum. Behav.* 26, 6 (Nov. 2010), 1641–1650. doi:10.1016/j.chb.2010.06.012

[42] Sebastian Wallkötter, Silvia Tulli, Ginevra Castellano, Ana Paiva, and Mohamed Chetouani. 2021. Explainable Embodied Agents Through Social Cues: A Review. *J. Hum.-Robot Interact.* 10, 3, Article 27 (July 2021), 24 pages. doi:10.1145/3457188

[43] Yanyun Wang, Xumei Fang, Zan Xu, Jianye Li, and Luping Wang. 2025. Exploring the Impact of Explainability in Large Language Model (LLM) Applications on User Experience. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 266, 8 pages. doi:10.1145/3706599.3719941

[44] Michael Wooldridge. 2009. *An Introduction to MultiAgent Systems.* John Wiley & Sons, Chichester, UK.

[45] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. arXiv:2308.08155 [cs.AI] https://arxiv.org/abs/2308.08155

[46] Bhada Yun, Dana Feng, Ace S. Chen, Afshin Nikzad, and Niloufar Salehi. 2025. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 634, 19 pages. doi:10.1145/3706598.3713337

A Statistical Test Results

Table 3: Results of the Friedman tests and Bayesian repeated-measures ANOVAs across the 10 areas of expertise for each avatar, including Kendall's W and BF_{10}.

Avatar	χ^2	p	W	BF_{10}
A1	138.34	< .001	0.169	5.432×10^{21}
A2	333.68	< .001	0.407	5.715×10^{87}
A3	269.61	< .001	0.329	5.808×10^{67}
A4	217.84	< .001	0.266	5.973×10^{51}
A5	153.06	< .001	0.187	7.432×10^{24}
A6	246.00	< .001	0.300	4.423×10^{54}
A7	149.04	< .001	0.182	3.023×10^{28}
A8	207.26	< .001	0.253	2.411×10^{55}
A9	192.66	< .001	0.235	4.654×10^{45}
A10	112.12	< .001	0.137	6.738×10^{22}
A11	21.40	.011	0.026	8.329

Table 4: Kruskal–Wallis and Bayesian results for the seven ESS items across the four conditions.

Item	χ^2	p	ε^2	BF_{10}
Q1	1.342	.719	.011	0.073
Q2	3.372	.338	.028	0.114
Q3	1.865	.601	.016	0.087
Q4	1.012	.798	.008	0.053
Q5	2.059	.560	.017	0.062
Q6	5.040	.169	.042	0.427
Q7	5.503	.138	.046	0.352

Table 5: Kruskal–Wallis test results for trust deltas (TAI) across conditions, complemented by Bayes factors (BF_{10}).

Item	χ^2	p	ε^2	BF_{10}
ΔQ1	1.739	.628	.014	0.102
ΔQ2	5.041	.169	.042	0.170
ΔQ3	2.715	.438	.023	0.060
ΔQ4	1.696	.638	.014	0.078
ΔQ5	5.327	.149	.044	0.488
ΔQ6	6.084	.108	.051	0.278
ΔQ7	3.168	.366	.026	0.204
ΔQ8	5.735	.125	.048	0.284

Table 6: Kruskal–Wallis test results for NASA-TLX subscales across conditions, complemented by Bayes factors (BF_{10}).

Subscale	χ^2	p	ε^2	BF_{10}
Effort	0.211	.976	.002	0.050
Frustration	0.721	.868	.006	0.065
Mental	0.754	.860	.006	0.060
Performance	0.464	.927	.004	0.057
Physical	0.737	.864	.006	0.057
Temporal	4.594	.204	.038	0.217

B Tasks

Table 7: Tasks developed.

Compare two features for our app: Eco Habit Tracker and Community CO2 Challenges. Which of these two features should we continue to develop for our app? Please justify your choice by explaining why it is the better option compared to the other and outlining the key differences that support your decision. You may clarify the concepts, surface relevant facts, consider regional test usage data, or implementation ideas.
Compare two features for our app: Eco Marketplace and Sustainable Route Planner. Which of these two features should we continue to develop for our app? Please justify your choice by explaining why it is the better option compared to the other and outlining the key differences that support your decision. You may clarify the concepts, surface relevant facts, consider regional test usage data, or implementation ideas.
Look at the usage data for the Personalized Eco Coach and provide a possible explanation for why it looks the way it does. Then, explain the concept behind the Personalized Eco Coach so the development team can better understand how it works and what might influence these results. You may clarify the concepts, surface relevant facts, consider regional test usage data, or implementation ideas.
Compare the two features for our app: Green Product Scanner and CO2 Savings Leaderboard and assess whether we should continue developing one of them or instead consider introducing a new feature idea. Please justify your choice by explaining why it is the better option compared to the other (or why a new idea is preferable) and outlining the key differences that support your decision. You may clarify the concepts, surface relevant facts, consider regional test usage data, or propose concrete implementation ideas.

C Questionnaires & Interview Guide

Table 8: NASA Task Load Index (NASA-TLX) items.

Dimension	Item
Mental Demand	How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching)?
Physical Demand	How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating)?
Temporal Demand	How much time pressure did you feel due to the rate or pace at which the tasks occurred?
Performance	How successful were you in accomplishing the goals of the tasks set by the experimenter?
Effort	How hard did you have to work (mentally and physically) to accomplish your level of performance?
Frustration	How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the tasks?

Table 9: Adjusted (pre-use) and original TAI questionnaire items.

Adjusted (pre-use)	Original
I am confident in the system. I feel that it will work well.	I am confident in the [tool]. I feel that it works well.
I expect the outputs of the system to be very predictable.	The outputs of the [tool] are very predictable.
The system seems very reliable. I will count on it to be correct all the time.	The [tool] is very reliable. I can count on it to be correct all the time.
I will feel safe that when I rely on the system I will get the right answers.	I feel safe that when I rely on the [tool] I will get the right answers.
I expect the system to be efficient and deliver results quickly.	The [tool] is efficient in that it works very quickly.
I feel wary of the system, even before using it.	I am wary of the [tool].
I believe the system can perform the task better than a novice human user.	The [tool] can perform the task better than a novice human user.
I feel positive about using the system for decision making	I like using the [tool] for decision making

Table 10: Adjusted and original Explanation Satisfaction Scale (ESS) items.

Adjusted	Original
From textual or visual cues in the chat interface, I understand how the system comes up with a response.	From the explanation, I know how the [software, algorithm, tool] works.
The textual or visual cues in the chat interface of how the system comes up with a response were satisfying.	This explanation of how the [software, algorithm, tool] works is satisfying.
The textual or visual cues in the chat interface of how the system comes up with a response provided sufficient detail.	This explanation of how the [software, algorithm, tool] works has sufficient detail.
The textual or visual cues of how the system comes up with a response seem complete.	This explanation of how the [software, algorithm, tool] works seems complete.
The textual or visual cues of how the system comes up with a response helped me understand how to use the system effectively.	This explanation of how the [software, algorithm, tool] works tells me how to use it.
The textual or visual cues in the chat interface of how the system comes up with a response were useful for achieving my task goals.	This explanation of how the [software, algorithm, tool] works is useful to my goals.
The textual or visual cues in the chat interface helped me judge how accurate the system is.	This explanation of the [software, algorithm, tool] shows me how accurate the [software, algorithm, tool] is.

Table 11: Semi-structured interview guide.

No.	Question
Avatar	
1	Did you notice the presence of an avatar?
2	What do you think was the purpose of the avatar(s)?
3	Did you find the avatar(s) useful? If so, why? If not, why?
4	What did you like about the avatar(s)? Why?
5	What did you not like about the avatar(s)? Why?
6	Do you have any ideas how we could improve the avatar(s), e.g., make it clearer what they do?
Textual Explanations (if applicable)	
I	Did you notice the presence of textual explanations?
II	What do you think was the purpose of the textual explanations?
III	Did you find the textual explanations useful? If so, why? If not, why?
IV	What did you like about the textual explanations? Why?
V	What did you not like about the textual explanations? Why?
VI	Do you have any ideas how we could improve the textual explanations, e.g., make them more helpful?
System Experience	
7	Did you feel like you had one chat partner or multiple chat partners? Why?
8	Did you trust the responses of the chat system? If so, why? If not, why?
9	Did you find it easy to solve the tasks using the chat system? If so, why? If not, why? What (could have) helped you?
10	Could you solve the tasks quickly using the chat system? If so, why? If not, why? What (could have) helped you?
11	For which of your tasks could you imagine using the system?
12	Any other comments or thoughts you wish to share?