SN Computer Science (2021) 2:194
https://doi.org/10.1007/s42979-021-00621-9

ORIGINAL RESEARCH

A Query Conundrum: The Mental Challenges of Using a Cognitive Assistant

Torsten Maier[1] · Saeed Abdullah[2] · Christopher McComb[3] · Jessica Menold[4]

Received: 12 November 2020 / Accepted: 25 March 2021

Abstract
Cognitive assistants use vocal interfaces and artificial intelligence to assist humans with complex tasks. While much research has focused on the application of these devices, a few studies have addressed how these devices affect the way humans work. To fill this gap, this research studied the effects of a cognitive assistant on mental workload, frustration, and effort. Participants worked with a Wizard-of-Oz style assistant and completed the Wisconsin Card-Sorting Task and engaged in a peripheral-detection task in a two-sample study that compared participants ($n=21$) who worked with the assistant to those who did not. Follow-up interviews were also completed. Results suggest that onboarding techniques, such as tutorials, are important for developing analogical trust before regular use. Additionally, results suggest that keeping the mental model of the CA clear, simple, and intuitive is important to reduce the mental effort that is required to account for the CA and interactions with it while working. Cognitive assistants offer a broad range of advantages but also have distinct challenges for users: primarily the lack of physical affordances that can be linked to functionality.

Keywords Cognitive assistant · Human–computer interaction · Mental workload · Trust · Wisconsin Card-Sorting Task

Introduction

Natural conversation through verbal and written language is becoming a common medium through which humans communicate with computers. Smart-speakers and chatbots already use natural language interfaces (NLI), and smart-speakers alone are present in 20% of households with Wi-Fi in the United States [1]. As these technologies progress, they will be able to support users during increasingly complex tasks using natural dialog, by helping users express their problems in a way that is compatible with technology, even with a limited user knowledge of the device [2]. The capabilities of cognitive assistants (technologies that combine NLIs and artificial intelligence, abbreviated CA) range from question–answer systems [3] to assisting in the design of Earth-orbiting satellites [4]. However, while more traditional physical tools require clear physical resources that the user can evaluate (e.g., lifting capacity, grip strength, etc.), CAs require mental resources that may be less clear to users and designers. Without a clear understanding of the HCI (human–computer interaction) fundamentals of CAs, it is impossible to identify which factors which factors may increase or decrease the mental effort required to engage with a CA. The objective of the current work was to provide fundamental information about how CAs affect human work through a generalizable task, providing insights that may guide designers to develop CAs with fewer obstacles for use, such as trust issues and mental overload. Specifically, this work addresses the research question *what mental challenges do users of CAs face and how do they affect the user*.

✉ Jessica Menold
jdm5407@psu.edu

Torsten Maier
torstennamaier@psu.edu

Saeed Abdullah
saeed@psu.edu

Christopher McComb
mccomb@psu.edu

[1] Industrial Engineering, The Pennsylvania State University, State College, USA

[2] Information and Science Technologies, The Pennsylvania State University, State College, USA

[3] Engineering Design, The Pennsylvania State University, State College, USA

[4] Engineering Design and Mechanical Engineering, The Pennsylvania State University, State College, USA

Published online: 04 April 2021

Related Work

The research in this paper spans a variety of topics from human–robot interaction (HRI) to social factors. First, we discuss relevant work from the HRI community and situate this work within the larger context of the field. HRI encapsulates the intersection of two components of this work: technology and social factors. In the following section, we discuss CA and mental workload in relation to existing work on HRI technology and social factors.

Human–Robot Interaction (HRI)

As robots continue to develop and become increasingly autonomous, their roles have transitioned from being operated on to working collaboratively and interactively with humans. This will require a new understanding of human–robot social factors, such as mental workload, emotional responses, and trust in a variety of contexts. For example, Visser and Parasuraman [5] studied static automation (robots that support humans continually) and adaptive automation (robots that support humans in critical situations). They found that adaptive automation improved participant's self-confidence, trust, and mental workload [5].

Mental workload is an important social characteristic, because it is strongly linked with performance [6–10]. Mental workload is defined as the cognitive cost required to achieve a certain degree of performance [11]. Measuring and tracking mental workload are a consistent challenge in HRI. Mental workload can be measured through self-report surveys, such as the *human–robot interaction workload measure* [12], or through psychophysiological responses, such as mean respiratory rate, respiratory rate variability, skin temperature [13], and eye tracking [14]. Transparency (i.e., the ability of the user to perceive the autonomous agent's abilities and develop an accurate mental model) has been linked to mental workload and situational awareness [15]. A mental model allows users to explain and predict the physical system they represent [16]. The properties of an object that define its possible uses are known as affordances. Technology with less physical components (e.g., smart-speakers) lacks physical affordances that help users generate mental models. Chen et al. found that increasing the transparency of UAV's autonomous capabilities significantly decreased subject's mental workload and increased situational awareness [15]. Additionally, accounting for the emotional state of the user can also lead to decreased mental workload in high-stress scenarios where collaboration with robots is crucial, such as rescue workers [17]. Affective (i.e., emotionally sensitive) robot functions can substantially enhance efficiency and effectiveness of cognitive workload through three core functions [17]:

1. Sliding autonomy: incorporates all intermediate levels of autonomy between tele-operation and full autonomy
2. Affective communication: the ability to recognize and understand utterances and affect, and have the ability to express them
3. Adaptive attitude: adaptation of the interaction to both the affective state of the user, and to the social relation between robot and user.

Accounting for the emotions of the user can also improve autonomous agents in areas other than mental workload. For example, emotion-sensitive natural language recognition has been shown to improve accuracy, which is important in high-stress situations, such as minimally invasive surgery controlled through natural language interfaces [18]. Moreover, robots expressing emotions and attention have improved social responses from participants [19].

Another social factor, trust, is an important factor in technology adoption [20]. Trust is defined as "the firm belief in the competence of an entity to act dependably, securely, and reliably within a specified context" [21]. The *Human–Robot Interaction Trust Scale* is the most common survey metric for measuring trust in HRI [22]. However, to properly measure dynamic trust (i.e., trust that changes over time), researchers must first determine which aspects of trust they are investigating and what type of trust. A literature review from Colquitt et al. distinguishes trustworthiness (the ability, benevolence, and integrity of a trustee) and trust propensity (a dispositional willingness to rely on others) from trust (the intention to accept vulnerability to a trustee based on positive expectations of his or her actions) [23]. In addition, Tan et al. split trust into dispositional (trust in other persons or machines upon initially encountering them, even if no interaction has yet taken place) and history-based trust (founded on interactions between the person and another person or machine) [24]. Finally, Lee and See define the characteristics that influence trust: analytic, analogical, and affective processes [25]. Analytical processes build trust through communicated knowledge, such as performance statistics. Analogical processes build trust through experience. Performance has been found to be the greatest influence of trust in analogical processes [26, 27]. The affective process builds trust through emotional connection.

Team dynamics also change as robots become team members as opposed to tools [28–30]. Successful team collaboration needs to consider how human–robot and robot–robot interactions affect team dynamics. For example, Tan et al. found that covert information exchanges

between robots in a human–robot team were less desirable than sharing information aloud [28]. Additionally, Luria et al. found that participants prefer agents that re-embody (move their social presence from body to body) rather than co-embody (move their social presence into a body that already contains another) [29]. Finally, shared cognition, or a shared understanding of the mental model between teammates," has been shown to be critical in effective human–robot collaboration [30].

Cognitive Assistants

Le and Watschinski define CAs as devices that "offer computational capabilities typically based on Natural Language Processing, Machine Learning, as well as reasoning chains operating on large amounts of data, enabling them to assist humans in cognitive processes" (p. 45) [31]. CAs enhance human capabilities, instead of replacing them. This is an important property that differentiates them from automation. The primary focus of research in this area has been on individual use-cases. For instance, StuA [3] and Duke [32] provide solutions in education. ADVICE [33] supports online e-commerce. A variety of CAs focus on aiding in the design process [34–37] and many CAs concentrate on medical applications [38–41]; for example, a personal health management that helps record observations in a self-care setting. Finally, CAs with specific functionality that span all fields are being developed, such as automated facilitators for virtual meetings [42, 43] and social networks [44].

While these studies provide insights about the utility of CAs for tasks and environments, understanding the HCI principles associated with CAs more broadly is understudied in comparison. This is problematic, because without this foundational knowledge to guide the field, standards and regulations cannot be developed knowledgably. These guidelines are especially important when it comes to the protection of vulnerable populations (i.e., people who are at risk of poor physical, psychological, and/or social health) [45]. NavCog, for example, is a CA that acts as a navigational aid to people with visual impairments [46].

Recent research within the HCI community has begun to address this gap through the development of standards to inform the design of CAs. Wolters et al. established guidelines for spoken dialog systems through the study of dementia patients [47]. Saad et al. formed metrics for measuring the "quality of experience" of CAs [48]. Finally, Tokadl developed design requirements for CAs used in space missions [49]. This research is informative, but nascent—more work is needed to understand the fundamentals of human and CA interactions.

Mental Workload and CAs

Mental workload is defined as the cognitive cost required to achieve a certain degree of performance [11]. Early theories proposed that humans draw upon a single undifferentiated pool of cognitive resources [50], while more recent theories proposed different mental resources for different types of tasks, known as the multiple resources model [51]. Mental workload is commonly measured through surveys and questionnaires such as the National Aeronautics and Space Administration Task Load Index (NASA-TLX) used in this study [11]. While there are many common mental workload questionnaires [52–56], the NASA TLX was selected, because it is a well-established instrument in engineering design research [57–59] and has been used in similar studies [60–63]. Additionally, mental workload can be measured via peripheral-detection tasks (PDT) [64–66]. PDTs typically require participants to respond to stimuli in a secondary task; the speed and accuracy of the response correspond to the mental workload of the participant [65]. Theoretically, as the mental workload of the primary task increases, researchers expect to see decreases in the speed and accuracy of responses on a secondary peripheral task.

A significant body of work has focused on the utility of CAs in automotive and aeronautics applications due to the link between mental workload and safety [6–10]. By decreasing the mental resources required for auxiliary tasks, such as monitoring vehicle gages, while driving/flying, more attention can be paid to the primary driving/flying task. Additionally, mental workload has been tied to performance [6–10]. For example, Svensson et al. found that airplane pilots' mental workload affected performance and information handling in 72 simulated low-level high-speed emissions [67]. Additionally, Lemoine et al. found decreased mental workload of air traffic controllers through collaboration with a CA [68]. Finally, in a similar study to the work presented in this paper, Kalnikaitė et al. tested the effect of a note-taking CA on mental workload and found that it improved recall although users found it more cognitively demanding [63].

In sum, HRI has demonstrated the importance of understanding different social and mental factors, such as trust, emotional response, and mental workload. CAs have the potential to be a revolutionary tool to enable humans, but are currently lacking research in HCI. This study aims to address this gap by investigating social and mental factors when using a CA through a generalizable task. The remainder of the paper is organized as follows: "Methods" section describes the methods and experimental design. "Results" section describes the analysis and results. "Discussion" section provides a discussion of the results. Finally, "Conclusion" section summarizes the findings and examines future work and limitations.

Methods

The current work aims to draw broader conclusions by investigating a generic task with follow-up interviews aimed at understanding how the CA affected users' mental workload among other cognitive factors. The Wisconsin Card-Sorting Task (WCST) [69] was selected as the generic task, because it is commonly used as a standard task for inducing mental strain, and has been employed in studies measuring differences in cognition that result from age [70, 71], diseases [72–74], and even alcohol consumption [75]. The combination of the NASA-TLX and peripheral-detection task (described in "Secondary speed and accuracy task (SSAT)" section) provides both, subjective and objective data, respectively. We also completed interviews to collect qualitative data on the participant's perception of the challenges faced during the primary task. The mixed methods used in this study aim to establish a complete picture of the phenomena investigated.

Study Design

Twenty-one participants took part in this study. All participants completed the Wisconsin Card-Sorting Task as their primary task, which was laid in front of them on a table. To measure mental workload, all participants also completed a Secondary Speed and Accuracy Test (SSAT) on the laptop next to them (a form of PDT). Participants were split into two conditions that modified how they completed the primary task. Ten participants were assigned to the control condition and completed the primary task without additional assistance. Eleven participants were assigned to the experimental condition and completed the task while working with a Wizard-of-Oz CA.

This CA was produced and manipulated using a Wizard-of-Oz prototyping technique, meaning that the researchers observed the participants from an adjacent room and manually controlled the CA to appear as if it was able to understand and communicate with the participants. Afterward completing the primary task, participants were given the NASA TLX to complete, as an additional measurement of mental workload. The NASA TLX is composed of six questions that query the user's perception of their mental, physical, and temporal demands, performance, and frustration.

Equipment and Facilities

The experiment was conducted in two connected rooms, the Observation Room and Experimental room (shown in Fig. 1). There was a one-way mirror between the two rooms which allowed the researchers in the Observation Room to directly observe participants in the Experimental Room. In the Experimental Room, there were several things placed on the table (displayed as gray rectangle), including the CA, a Bluetooth speaker, a laptop, and the Wisconsin Card-Sorting cards. In the Observation room, there was an experimenter's laptop to run the CA. Due to the sound insulation between the two rooms, an audio feed was established via two cellular devices, connected between the two rooms.

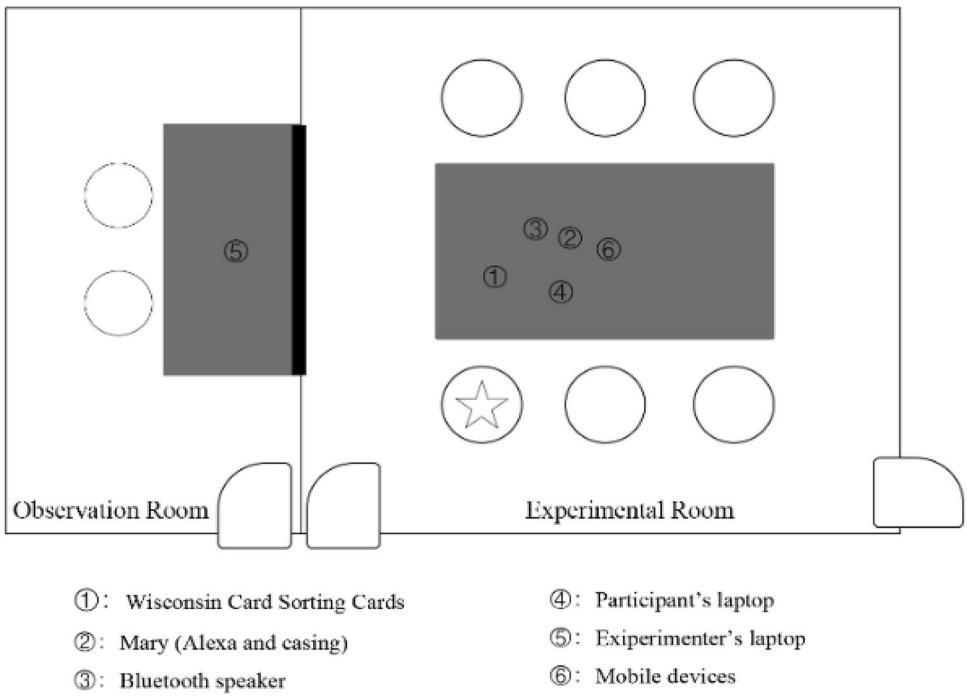

Fig. 1 Overhead view of the experimental setup

Participants

Participants were recruited through a purposeful snowball sampling technique (a sampling technique where existing participants recruit future participants) [76] and signed up electronically in accordance with university Internal Review Board procedures. A total of 21 participants were recruited from the Departments of Information Sciences and Technology, Computer Science, and Industrial Engineering at a public university. Participants were randomly assigned to two groups. In the experimental group, participants completed the task with the CA, "Mary" ($n=11$), and in the control condition, participants completed the task without "Mary" ($n=10$). Participants were between 20 and 29 years of age; 12 participants were male and 9 were female. Following the experiment, all participants were invited to take part in a follow-up interview. In total, 8 participants accepted this invitation.

Experimental Design

To answer our research question, *what mental challenges do users of CAs face and how do they affect the user,* a controlled study was conducted. To measure mental workload, we recorded reaction time to the SSAT, their SSAT accuracy (calculated by number of correct responses divided by the total possible), and the NASA TLX Scores. Participants' interaction frequency (how many times participants asked Mary for help) and participants' performance in the WCST (calculated by number of correct cards divided by the total possible cards) were also recorded.

A full script for the study is provided in Appendix A and summarized here. At the start of the study, the researcher provided a brief introduction to participants and obtained informed consent, according to Internal Review Board (IRB) protocol. Participants then completed a demographics survey. Once they finished the survey, experimental details were carefully explained by the researcher, including how to do the WCST, how to do the SSAT, and when to start and stop. Additionally, participants were informed that the CA was roughly 95% accurate to provide a basis of analytical trust (trust based on analysis or logical reasoning [25]) for the experiment. The exact phrasing used by the researcher was (the full study script can be found in Appendix B):

> *"You may ask the Assistant at any point for a suggestion and she will suggest a rule that you should try by analyzing your past turns. In prior testing, the Assistant was found to be roughly 95% accurate."*

To help the participants better understand the tasks, they were required to practice once under the supervision of the researcher (five cards in total were used for the practice trial). Participants were free to ask any questions during this period. Without further questions, the researcher would leave the experimental room and the primary experiment began. While participants were doing the task, all the behaviors were observed by researchers in the Observation Room. The WCST sounds and the CA's suggestions were all directed by the researchers, and were the only feedback provided to the participant. Once participants completed the WCST, the experimenter came back to the room and allowed participants to finish the NASA TLX [57–59] (a survey to help measure their mental workload). Finally, the experimenter would debrief the participants on the purpose of the study and recap the IRB disclaimer. No qualitative data (e.g., observational behaviors or spoken dialog) were recorded during the primary task; however, follow-up interviews were completed with 8 participants to provide a deeper understanding of their actions during the study.

Wisconsin Card-Sorting Task (WCST)

In short, in the WCST, participants must sort cards according to the active sorting criteria that must be inferred by the participant throughout the task [69]. There are four potential sorting criteria, including the color of the symbols, the shape of the symbols, the number of the shapes on each card, and the background (as seen in Fig. 2). The WCST can be altered based on the card deck size, order of the cards, and the sorting strategies [77]. Our WCST included 31 cards, shown in a fixed order for each participant, and 4 sorting criteria.

There are 4 base cards that do not change. The participant draws a card from the draw pile and places it in one of the four placement areas which correspond to the above base cards. The participant then hears a correct or incorrect sound played over the speaker and then places that card in the discard pile and repeats the process. This sound is the only feedback the participants receive to independently infer the correct matching rule.

Most participants complete the task through trial and error. The classification rule changes every five cards (31 cards in total, the final rule lasts for 6 cards), introducing

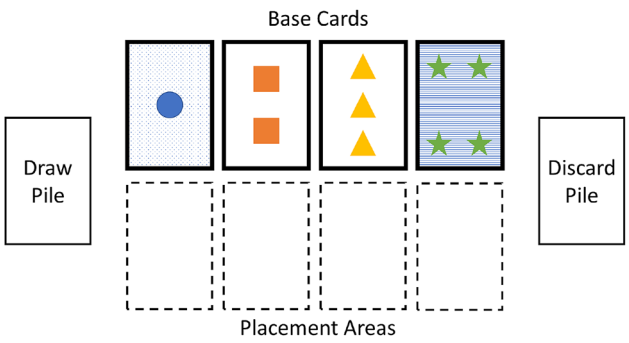

Fig. 2 Wisconsin card-sorting task

complexity into the task. For this task, performance was measured by dividing the total number of correctly placed cards by the total number of cards.

Secondary Speed and Accuracy Task (SSAT)

Peripheral Detection Tasks (PDT) use the performance of a peripherally located secondary task to estimate the mental workload of the centrally located primary task [65]. PDTs rely on signal detection theory to classify user responses into hits, misses, false alarms, and correct rejections [78]. PDTs have seen use in fields ranging from automotive [64, 79, 80] to interface design [81] to medicine [82].

For this study, the authors developed their own PDT using a peripherally located laptop. The PDT was implemented to measure mental workload during the completion of the primary task [64, 65, 79, 81]. Due to the hierarchal nature of attention, even simple secondary tasks can be an effective measure for attention [65]; for example, see [64, 79, 81]. The interface for the SSAT was developed using PsyToolkit [83, 84], a website that provides a toolkit for demonstrating, programming, and running cognitive-psychological experiments and surveys. During this test, left and right arrows appear on the monitor. Whenever an arrow appears, the participant needs to strike the arrow keys on the keyboard corresponding to the arrow presented on the screen. During this test, hits, incorrect hits, and misses are recorded. The reaction time and accuracy measures are used to estimate the participants' mental workload.

Cognitive Assistant (Mary)

To allow for flexibility in the experimental design, a Wizard-of-Oz approach was used to simulate an autonomous CA [85, 86]. Wizard-of-Oz style experimental designs have been used in previous CA-related work, including speech user interfaces [87] and instructional bots [85]. However, they have seen relatively little use for the design of intelligent systems in the mechanical design community. An instance of this kind of work, Jou et al. [88] used a Wizard-of-Oz approach to demonstrate water reduction through an "autonomous" faucet [88]. During the initial development of this study, the experimenters were able to make quick modifications to the simulated CA using the Wizard-of-Oz approach. Additionally, this was a low-cost method that enabled the research team to evaluate human-CA interactions without the technological burden of constructing a complete CA system (Fig. 3).

An Amazon Echo in the Experimental Room was connected to a laptop in the Observation Room via a Bluetooth connection and, when prompted by the participant for suggestions, corresponding voice lines were chosen by the experimenter in the Observation Room. The Amazon Echo was shrouded by a 3D printed casing as to not bias the participants by the presence of the Echo. Mary could be asked by the participants to suggest their next rule guess, to which the CA would provide suggestions by saying "try color", "try background", "try shape", or "try number". For this study, the CA always provided participants with the correct answer. However, participants were not made aware that the CA would always be correct. Participants had to assess the correctness of the CA on their own throughout the experiment. This is further discussed during the interview section of the paper. Yet, future studies investigating the effect of trust on usability in CAs could vary the accuracy of the CA used here. The number of requested suggestions by a participant throughout an experiment was recorded (when working with a CA). Participants requested 6 ± 3 suggestions per experiment. An example participant can be seen completing the experiment in Fig. 4.

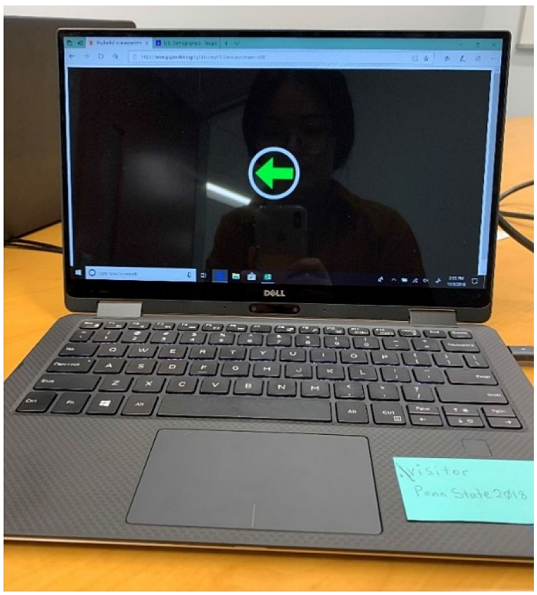

Fig. 3 Secondary speed and accuracy test

Interview Methods

To gain a deeper understanding of the phenomena discussed in the prior sections, eight follow-up interviews with participants from the study were completed. Four interviewees completed the study with the CA and four interviewees completed the study without. Participants were recruited using the email address provided by the participant during the initial study. The semi-structured interviews lasted approximately 10 min. Questions focused on the perceptions of the primary task and the CA (if they were in the group that interacted with it).

Fig. 4 Example participant doing the task

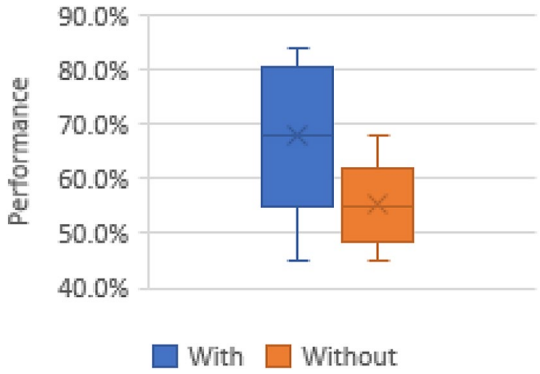

Fig. 5 Box plot showing greater performance with the CA than without

Results

For all metrics, an independent unpaired t test was performed. Normality of the data was confirmed by checking that the skewness and Kurtosis values were between − 2 and + 2 (produced by the Descriptive Statistics test in Excel's Data Analysis ToolPak). Homogeneity of variances checked using Levene's test (produced by a Single Factor ANOVA test in Excel's Data Analysis ToolPak). Outliers were greater than three standard deviations from the mean. No outliers were found. Significance values greater than 0.1 were considered insignificant, values between 0.1 and 0.05 were investigated for both practical and statistical significance, and values less than 0.05 were considered significant. The threshold of 0.05 to 0.1 for further investigation was based on previous work, indicating that other statistical indicators (such as effect size) are better suited for weighing the implications of results given the arbitrary nature of the 0.05 threshold [89, 90]. Effect sizes less than 0.4 were considered small, sizes between 0.4 and 0.7 were considered moderate, and sizes greater than 0.7 were considered large. This section examines the participant performance on the WCST, performance on the SSAT, and responses to the NASA TLX.

Wisconsin Card-Sorting Task Performance

Performance on the WCST was measured by dividing the total number of correct cards by the total number of possible cards (31), producing a percentage of correct cards indicating the performance of the participant. A box plot of performance is provided in Fig. 5. The middle line indicates the median, the X represents the mean, and the lower and upper whiskers extend to the first and third quartile, respectively. This study found that participants using the CA had statistically significantly higher performance scores ($67.7 \pm 8.6\%$) compared to participants without the CA ($55.5 \pm 5.4\%$), $p = 0.023$. Additionally, the effect size was large, $d = 0.96$. In other words, the CA produced the desired effect of increasing the performance of the participants using it.

It may be expected that higher rates of CA usage would correlate with higher performance. However, a linear regression analysis using the number of suggestions (independent variable) and the performance (dependent variable) found no significant relationship ($r^2 = 0.17$, $p = 0.21$, performance $= 0.5536 + 0.0203x$). This may imply that within the card-sorting task, participants scaled their use of the CA (i.e., participants already doing well did not feel the need to use the CA unlike participants doing poorly who relied on the CA more heavily). In future investigations, the motives of the participants should be studied as well.

Secondary Speed and Accuracy Task

The secondary speed and accuracy task are composed of two metrics: reaction time and accuracy. Reaction time is measured from the moment that the arrow appears on the screen to the moment that the participant selects an arrow on the keyboard. The difference in reaction time between participants with the CA (1878 ± 209 ms) and without (1968 ± 539 ms) was not significant, $p = 0.73$. The effect size was also small, $d = 0.16$. This indicates that the CA did not impact how quickly participants were able to spot and react to the secondary task. This suggests that CAs do not induce more cognitive load in users completing a simple task, as compared to the users working without a CA.

Table 1 NASA TLX average and standard deviations

	Average (with)	Average (without)	Stdev (with)	Stdev (without)
Mental demand	10.64	10.20	5.26	4.44
Physical demand	4.73	5.60	4.33	5.36
Temporal demand	8.09	9.10	4.09	3.75
Performance	8.64	9.60	3.74	2.41
Effort	9.64	11.80	4.39	3.46
Frustration	3.82	6.00	2.94	2.83
Average	7.59	8.72	2.27	1.96

Table 2 NASA TLX significance and effect size

	Significance	Cohen's *d*
Mental demand	0.853	0.083
Physical demand	0.665	0.202
Temporal demand	0.583	0.247
Performance	0.559	0.258
Effort	0.262	**0.493**
Frustration	**0.090**	*0.743*
Average	0.262	**0.495**

Significance: bold for less than 0.1. Effect size: moderate effect between 0.4 and 0.7 (bold) and large effect greater than 0.7 (italics)

Accuracy is calculated by dividing the number of correct responses by the total number of responses. Higher accuracy scores indicate that participants could focus more on the secondary task, because the primary task was less cognitively demanding. The difference in accuracy between participants with the CA ($94.4 \pm 5.7\%$) and without ($90.8 \pm 7.9\%$) was not significant, $p = 0.36$. However, the effect size was moderate, $d = 0.43$, indicating that accuracy is trending toward significance pending increased sample sizes. This may be evidence that participants using the CA experienced lower cognitive load. Future studies will continue to explore this trend.

NASA TLX

The NASA TLX is composed of six metrics: mental demand, physical, temporal demand, performance, effort, and frustration. The averages and standard deviations are provided in Table 1. The significance and effect size are provided in Table 2. This study found that participants using the CA had statistically significantly lower frustration scores (3.82 ± 1.97) compared to participants without the CA (6.00 ± 2.02), $p = 0.090$. Additionally, the effect size was large, $d = 0.74$. In other words, we found evidence that CAs reduce the amount of frustration users' experience when engaging in repetitive tasks; the large effect size signifies that this finding is robust and suggests CAs may effectively reduce frustration during simple tasks.

Table 3 The number of suggestions vs. NASA TLX scores

	Significance	r^2
Mental demand	0.657	0.023
Physical demand	0.520	0.047
Temporal demand	**0.076**	**0.308**
Performance	0.989	0.000
Effort	0.486	0.055
Frustration	0.551	0.041

Significance: bold for less than 0.1. Effect size: moderate effect between 0.4 and 0.7 are bold

Suggestions vs. NASA TLX Scores

Additional analysis was completed to investigate the relationship between the number of suggestions requested by the participant (independent variable) and the NASA TLX scores (dependent variable). From Table 3, only Temporal Demand trended toward significance ($p = 0.076$) as well as providing explanatory value ($r^2 = 0.308$). The regression equation was Temporal Demand $= 13.44 - 0.879x$ (the number of suggestions requested). This suggests that the more the CA assisted an individual, the less hurried the participant felt. This is particularly interesting, because there was not a time limit imposed on participants during the experiment. However, this trend should be further studied in the following research.

Interviews

Several key insights were derived from an inductive analysis of the eight interviews. These insights can be categorized into three major themes: the necessity for trust, overcoming getting stuck, and how the CA design can both help and hurt.

Interviewees Without the CA

All four interviewees who performed the task without the CA discussed similar methods for completing the WCST and overcoming challenges in their interviews. These

participants followed a "guess and check" method where a new rule would be guessed until a correct rule was found. This rule was then followed until it became incorrect, and then, the "guess and check" process would begin again. When these participants got stuck (repeated incorrect guesses), they did not alter or devise a new strategy, but instead continued "guessing and checking" until eventually they found another correct rule. These longer periods of confusion led to frustration among the participants. Suggestions for what would have helped them in the task included being able to see their previous card placements and tips on where to place their next card.

The Necessity for Trust

A central tenant that all four interviewees who worked with the CA discussed was the idea of trust.

> Like it would give me options, I did not really know if it was something I should go for or not. I still had to rationale it out in my head and then make the next decision based on my internal system. I didn't know whether I could trust it or not. -Participant 2

Another participant also mentioned:

> 'I think if she was going to be wrong some of the times I think her answer would have been probably similar to guessing the card myself so if I'm not sure that she's right or wrong I at least know the probably of the guesses that I'm actually doing right now.' -Participant 3

Interviewees described a natural progression from infrequent use to dependent frequent use as they gained trust in the CA. Like a learning curve, a trust curve model should be considered when designing CAs to account for this progression. Potential trust curves have been presented in [91].

It should also be noted that this trust progression is affected by the performance of the system. Low system performance may result in slowed trust-building, no trust-building, or even an erosion of trust altogether. While high system performance may result in accelerated trust-building.

Lee and See identified three primary components of trust in automation: analytical, analogical, and affective trust [92]. Participants were told at the beginning of the experiment that the CA was roughly 95% accurate, thus building analytical trust (see Appendix A). However, participants needed to build a basis of analogical trust by repeatedly experiencing the CA accurately predicting the correct card placement. Once this trust had been established, participants felt comfortable calling on the CA more frequently. Although, the effects of affective trust were not observed here, this work shows the importance of developing all three forms of trust to ensure user interaction, regardless of system performance.

Overcoming Getting Stuck

The first stage of the trust progression started, for most interviewees, by getting stuck. Interviewees began the experiment with the mindset that they had to complete the task primarily by themselves. However, when they began to run into roadblocks, the CA became the best alternative to turn to. By seeing the CA accurately predict the correct card placement repeatedly, interviewees started seeing the CA as an option other than just a last resort.

> At first, I really wanted to do it all myself and only get help when I need it and then it was like ok, that doesn't have to be a thing. I can ask Mary as many times as I want. -Participant 1
> I started using her only after I found out that, I was stuck with a few cards and I tested out once or twice, I knew that she was going to be right all the time, so that's when I was dependent on her when I got stuck. -Participant 3

On the other hand, interviewees who did not work with the CA had no alternative options.

> I think I just continued guessing a bunch of different cards until something hit. -Participant 8

They resorted to repeating the same actions with the hope of a different outcome, creating a feeling of frustration. This may indicate that even if the performance of a CA is like that of the human, merely providing an alternative may mitigate participant frustration. However, this hypothesis requires further testing.

CA Design Can Help and Hurt

All four interviewees that worked with the CA agreed that it was helpful. However, interviewees also noted that at times, the CA also increased their mental workload. Interviewees described a trade-off where the CA resulted in a constant additional increase in workload, but also mitigated the spike in workload normally produced by periods of confusion (i.e., the moments when the rules would change in the primary task).

> It was another thing to do but it took the load off from "I don't remember what the last card was", "I don't know what the pattern has been" I'll just ask Mary. -Participant 1

In effect, the CA flattened the workload of the participants at the cost of elevating the base workload. This

trade-off should be considered when designing CAs and explored further in future work. Such studies are critical for the future mindful development of Cas, because if the baseline or spikes are too high, it is likely that users will reject the system.

Discussion

Performance vs. Mental Workload

The goal of this work was to explore the effect of CAs on mental workload in a card-sorting task. Since the CA always provided the participants with the correct answer, it was expected that those participants using it would perform significantly better at the WCST. However, it is noteworthy that this increased performance did not come at a cost to mental workload, as seen from the SSAT and NASA TLX scores. Our results suggest that for simple repetitive tasks, CAs could increase performance without adding additional mental strain to workers. It is possible that there was no difference in mental stress in this experiment, because the CA and the primary task required different mental resources. Based on Wicken's Multiple Resource model, auditory (CA) and visual (WCST) require separate mental resource pools and would, thus, have less of an impact on each other than two audio or two visual tasks [51]. Our findings support the growing body of literature surrounding the design of CAs and the effect design decisions may have on mental workload. For example, Strayer et al. [7] found that CAs increased mental workload in driving tasks, yet Brookhuis et al. [6] found that CAs decreased mental workload in a similar task; these conflicting findings point to the effect the design and modality of the CA has on mental workload.

Frustration

The NASA-TLX frustration scores showed that users using the CA reported being less frustrated as compared to the control group. CA technology could be further improved by prompting the device to assist the user, not only when the user requests, but also upon behavioral cues that may indicate frustration. This is known as adaptive affective computing, an approach which infers the users' emotions and adapts the graphical interface to counter user frustration [93]. Real-time or just-in-time interventions via constant data collection could prove incredibly beneficial [94] as intervention techniques could include detecting users' emotional state and providing a real-time strategy to decrease users' stress/frustration.

Trust

Qualitative data from semi-structured interviewees most participants cited a lack of trust in the CA as the key inhibitor preventing them from using it at each opportunity. There are three types of trust (analogical, analytical, and affective): analogical trust is built through experience, analytical trust is built through shared information, and affective trust is built through emotional attachment [25]. In the interviews, participants cited analogical trust as the primary means of building trust in the CA.

> So, once I started to build confidence that Mary knew what she was talking about, I think I started asking her more often. -Participant 1

This trust-building process was often catalyzed by the presence of an obstacle. Meaning that the presence of the obstacle forced users to rely on the CA, showing them that the CA could be trusted. This phenomenon is also seen in human-to-human relationships; for example, team-building exercises use challenges to build collaboration and trust amongst team members [95]. However analytical trust was also built pre-task when the participants were told the CA was roughly ~ 95% accurate and affective trust was built through the name and voice of the CA. This may imply that while they can build a strong foundation of trust through affective and analytic trust, they will still need to go through an introductory phase of analogical trust-building before they can expect users to interact with their CAs as expected. This could be combatted with tutorials which allow the users to see the CA working with a 100% success rate on predefined examples. Future studies will investigate this phenomenon further.

Mental Workload

Finally, the quantitative and qualitative results describe a complex effect of CAs on participants' mental workload. The quantitative results found no significant difference in the secondary task between the control and CA groups. However, in the post-task interviews, participants provided a more nuanced explanation, describing how, at times, the CA hurt or helped their mental workload. Participants described the CA as "another thing to think about," increasing their perceived mental workload for the portion of the session where they were not actively engaged with the CA. Conversely, participants also stated that they felt a decrease in their perceived mental workload during the portion of the session where they engaged with the CA to assist them when the primary task confused them. This occurred most often when the sorting rule would change in the WCST. Because the CA helped and hurt participant's perceived

mental workload, this may explain why the quantitative results found no significant differences.

This trade-off in mental workload is important for CA designers to consider, because a consistent increased level of mental workload may be overburdening for some jobs. Additionally, unlike physical items or even graphical interfaces where connections and interactions can be clearly displayed, CAs rely entirely on audible signals, requiring users to have a clear mental model of the expected interactions. Thus, it is incredibly important to combat these problems by keeping the mental model of the CA clear, simple, and intuitive so little mental work is required to account for the CA and its interactions while working.

The properties of an object that define its possible uses are known as affordances. Technology with less physical components (e.g., CAs like smart-speakers) lack physical affordances that help users generate mental models. The functionality and limitations of CAs lie in their ability to understand and execute the user's intentions through natural language. Invisible affordances like these that exist but are in a system too large and ambiguous to easily discern must be intentionally revealed. Prior research in the field has presented similar findings in different niches. Tolmie et al. proposed novel approaches to handling technological interruptions through "making the grounds of disturbance visible and available to practical reasoning" [96]. Chen et al. found that increasing the transparency of a UAV's autonomous capabilities, subject's mental workload was significantly decreased and situational awareness was increased [15]. Along with building trust, the aforementioned tutorials could be used to clarify the CA's capabilities and controls (normally perceived through affordances).

Limitations

In both a limitation of this study and an area for future work, participants in this study were not asked to complete a personality assessment nor where they queried for their previous experiences with AI/CA technology. Personality types, for example, could have been investigated for any links to emotional responses such as frustration. Both avenues will be investigated in future work. Additionally, future work will consider a larger and more diverse sample population as most participants were engineering/information science and technology (IST) students. Participants from different fields (e.g., psychology, biology, history, etc.) and different experience levels (i.e., professionals) may have different perspectives. Furthermore, the number of participants in the control and task groups, respectively, is too few to be generalized broadly. Thus, further work on this research question is required to generalize these findings to a broader degree. Finally, the CA used in this study was 100% accurate with its suggestions to the participants. Participants were explicitly informed that in early testing, the CA demonstrated 95% accuracy with suggestions. It is possible some participants may have realized the CA was 100% accurate which may have resulted in an accelerated trust relationship between the CA and the participant. We note, however, the investigation of errors and their effect on the mental factors influenced by working with CAs was outside the scope of this work and is being fully investigated in the following studies.

Conclusion

Our research study aimed to answer the following research question: *what mental challenges do users of CAs face and how do they affect the user*. Guided by a mixed-methods study design, this study investigated the effect of CAs on mental workload, frustration, and trust. Findings indicate that CAs can improve the performance of a simple task and that CAs have a complex effect on mental workload. Additionally, through the analysis of follow-up interviews, it was recommended that tutorials or other onboarding techniques should be further investigated for Cas, because they provide a means of developing analogical trust and intentionally revealing invisible affordances before regular usage. Finally, it is important that CAs have a clear, simple, and intuitive structure, so that minimal mental work is needed when considering how to use the CA.

Appendices

Appendix A: Post-Task Interview Questions

1. How difficult did you find the card-sorting task?
2. How frustrating did you find the task?
3. What methods did you use to solve the card-sorting task?
4. How did you overcome getting stuck?
5. Did you work with the CA?
    a. If yes
        i. How and when did you intend to use the CA?
        ii. Do you feel the CA helped, why or why not?
        iii. Do you feel the CA added to your workload?
    b. If no
        i. What would have helped you complete the card-sorting task?

Appendix B: Study Script

All

Thank you for agreeing to participate in our study. All of the information collected is anonymous. You have the right to withdraw your participation, or any/all of your data without giving a reason and retrospectively.

Today you will take a demographic survey, then we will cover the details you need to complete the study, then you will take a post-task survey, and finally, we will wrap up with some more information.

Before we begin, please complete the U.S. Demographics Survey provided on the laptop before you.

Wait for completion of survey.

Thank you. Today you AND YOUR ASSISTANT will be completing the Wisconsin Card-Sorting Task. In this task, you have 4 base cards.

Researcher points to the four base cards laying in front of the participant.

You will need to select a card from the deck.

Researcher points to the deck laying in front of the participant.

And place that card in front of one of the 4 base cards that you believe matches the current rule. The possible rules are the color of the symbols (blue, orange, yellow, or green), the symbols themselves (circle, square, triangle, or star), the number of the symbols (1, 2, 3, or 4), or the style of background (striped, white, or dotted). For the white background rule, you may place it on either white background card. Once a card is placed, you will hear a sound for correct or incorrect.

Regardless of correct or incorrect, you will then place that card on the discard pile.

Researcher points to the discard pile area in front of the participant.

You will repeat this process until all cards have been moved to the discard pile. As you go along, the rules you are matching will change.

Do you have any questions?

Answer questions if necessary.

For participants using the CA

To help you with this task, you will be working on a team with our own personal assistant. Assistant, please introduce yourself

Researcher 2 plays introduction on assistant

You may ask the Assistant at any point for a suggestion and she will suggest a rule that you should try by analyzing your past turns. In prior testing, the Assistant was found to be roughly 95% accurate

Do you have any questions about how to interact with the Assistant or what she can be used for?

Answer questions if necessary

All

While you complete the Wisconsin Card-Sorting Task, you will be doing a secondary speed and accuracy test. Throughout the task, you will see either a left or right arrow appear. When you see the arrow appear, please select the corresponding left or right arrow on the keyboard. The secondary task will continue to run after you have completed the Wisconsin Card-Sorting Task but please stop doing the secondary task when you discard your last card. Let the secondary task run itself out.

Please practice the secondary task until you feel comfortable.

Researcher 1 shows a sample secondary task.

Now we are going to go through a quick practice trial.

Run through practice trial.

Do you have any questions?

Answer questions if necessary.

I will now leave the room. Once I have closed the door, you may begin. Once you've completed the task, I will return, and we will debrief.

Researcher 1 leaves and participant completes the experiment.

Thank you for your participation. Please complete this post-task survey.

The aim of this study is to measure the amount of mental strain produced by our Assistant. Half of the participants will be doing the task with the Assistant and half will be doing the task without her. We measured how well you did in the Wisconsin Card-Sorting Task, your reaction time and accuracy in the secondary task, how often you interacted with the Assistant, and your perceived workload from the post-task questionnaire.

All of the information collected is anonymous. You have the right to withdraw your participation, or any/all of your data without giving a reason and retrospectively. At any point, you make contact with the principal investigators. Here is their contact information.

Provide contact information.

Availability of data and materials https://github.com/torstennamaier/Query_Conundrum.

Code availability https://github.com/torstennamaier/Query_Conundrum.

Declarations

Conflict of interest The authors declare no conflict of interest.

References

1. Bernard Z. Markets insider. 2018. https://markets.businessinsider.com/news/stocks/smart-speakers-are-taking-off-with-consumers-charts-2018-4-1021723081.
2. Castellani S, Grasso A, O'Neill J, Roulland F. Designing technology as an embedded resource for troubleshooting. Comput Support Coop Work. 2009. https://doi.org/10.1007/s10606-008-9088-1.
3. Lodhi P, Mishra O, Jain S, Bajaj V. StuA: an intelligent student assistant. IJIMAI J. 2018:1–9.
4. Bang H, Martin A, Prat A, Selva D. Daphne: an intelligent assistant for architecting earth observing satellite systems. AIAA Conf. Proc. 2018. https://doi.org/10.2514/6.2018-1366.
5. De Visser E, Parasuraman R. Adaptive aiding of human-robot teaming: effects of imperfect automation on performance, trust, and workload. J Cogn Eng Decis Mak. 2011. https://doi.org/10.1177/1555343411410160.
6. Brookhuis KA, Hoedemaeker M, van Arem B, van Driel CJG, Hof T. Driving with a congestion assistant; mental workload and acceptance. Appl Ergon. 2008;40:1019–25. https://doi.org/10.1016/j.apergo.2008.06.010.
7. Strayer DL, Cooper JM, Turrill J, Coleman JR, Hopman RJ. The smartphone and the driver's cognitive workload: a comparison of apple, google, and microsoft's intelligent personal assistants. Can J Exp Psychol. 2017;71:93–110.
8. Estes S, Helleberg J, Long K, Pollack M, Quezada M. Guidelines for speech interactions between pilot & cognitive assistant. In: ICNS 2018—Integr. Commun. Navig. Surveill. Conf. 2018. p. 1–23. https://doi.org/10.1109/ICNSURV.2018.8384965.
9. Pitchammal R, Sadda V. Making the mission computer intelligent-a step ahead. Def Sci J. 2013;63:174–80. https://doi.org/10.14429/dsj.63.4260.
10. Wilkins SA. Examining head-down time in transportation: case study in single-pilot general aviation operations, Transp. Res. Rec. J. Transp. Res. Board. 2018:036119811877652. https://doi.org/10.1177/0361198118776521.
11. Hart SG, Staveland LE. Development of NASA-TLX (Task Load Index). Adv Psychol. 1988;52:139–83. https://doi.org/10.1016/S0166-4115(08)62386-9.
12. Yagoda RE. Development of the Human Robot Interaction Workload Measurement Tool (HRI-WM). Proc Hum Factors Ergon Soc Annu Meet. 2010. https://doi.org/10.1177/154193121005400408.
13. Novak D, Mihelj M, Munih M. Psychophysiological responses to different levels of cognitive and physical workload in haptic interaction. Robotica. 2011. https://doi.org/10.1017/S0263574710000184.
14. Buettner R. Cognitive workload of humans using artificial intelligence systems: Towards objective measurement applying eye-tracking technology. In: Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), 2013. https://doi.org/10.1007/978-3-642-40942-4-4.
15. Chen T, Campbell D, Gonzalez F, Coppin G. The effect of autonomy transparency in human-robot interactions: a preliminary study on operator cognitive workload and situation awareness in multiple heterogeneous UAV management. In: Australas. Conf. Robot. Autom. ACRA, 2014.
16. Greca IM, Moreira MA. Mental models, conceptual models, and modelling. Int J Sci Educ. 2000. https://doi.org/10.1080/095006900289976.
17. Looije R, Neerincx M, Kruijff GJM. Affective collaborative robots for safety & crisis management in the field. In: Intell. Hum. Comput. Syst. Cris. Response Manag. ISCRAM 2007 Acad. Proc. Pap. 2007.
18. Schuller B, Rigoll G, Can S, Feussner H. Emotion sensitive speech control for human-robot interaction in minimal invasive surgery. In: Proc. 17th IEEE Int. Symp. Robot Hum. Interact. Commun. RO-MAN, 2008. https://doi.org/10.1109/ROMAN.2008.4600708.
19. Bruce A, Nourbakhsh I, Simmons R. The role of expressiveness and attention in human-robot interaction, In: Proc. - IEEE Int. Conf. Robot. Autom. 2002. https://doi.org/10.1109/robot.2002.1014396.
20. Bahmanziari T, Pearson JM, Crosby L. Is trust important in technology adoption? A policy capturing approach. J Comput Inf Syst. 2003. https://doi.org/10.1080/08874417.2003.11647533.
21. Grandison T, Sloman M. A survey of trust in internet applications. IEEE Commun Surv Tutorials. 2009;3:2–16. https://doi.org/10.1109/comst.2000.5340804.
22. Yagoda RE, Gillan DJ. You want me to trust a ROBOT? The development of a human–robot interaction trust scale. Int J Soc Robot. 2012. https://doi.org/10.1007/s12369-012-0144-0.
23. Colquitt JA, Scott BA, LePine JA. Trust, trustworthiness, and trust propensity: a meta-analytic test of their unique relationships with risk taking and job performance. J Appl Psychol. 2007. https://doi.org/10.1037/0021-9010.92.4.909.
24. Merritt SM, Ilgen DR. Not all trust is created equal: dispositional and history-based trust in human-automation interactions. Hum Factors. 2008. https://doi.org/10.1518/001872008X288574.
25. Lee JD, See KA. Trust in automation: designing for appropriate reliance. Hum Factors. 2004. https://doi.org/10.1518/hfes.46.1.50_30392.
26. Hancock PA, Billings DR, Schaefer KE, Chen JYC, De Visser EJ, Parasuraman R. A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors. 2011. https://doi.org/10.1177/0018720811417254.
27. Salem M, Lakatos G, Amirabdollahian F, Dautenhahn K. Would you trust a (faulty) robot?: Effects of error, task type and personality on human-robot cooperation and trust. ACM/IEEE Int Conf Human-Robot Interact. 2015. https://doi.org/10.1145/2696454.2696497.
28. Tan XZ, Reig S, Carter EJ, Steinfeld A. From one to another: how robot-robot interaction affects users' perceptions following a transition between robots. ACM/IEEE Int Conf Human-Robot Interact. 2019. https://doi.org/10.1109/HRI.2019.8673304.
29. Luria M, Reig S, Tan XZ, Steinfeld A, Forlizzi J, Zimmerman J. Re-embodiment and co-embodiment: Exploration of social presence for robots and conversational agents. In: DIS 2019 - Proc. 2019 ACM Des. Interact. Syst. Conf., 2019. https://doi.org/10.1145/3322276.3322340.
30. Demir M, McNeese NJ, Cooke NJ. Understanding human-robot teams in light of all-human teams: aspects of team interaction and shared cognition. Int J Hum Comput Stud. 2020. https://doi.org/10.1016/j.ijhcs.2020.102436.
31. Le N-T, Wartschinski L. A Cognitive Assistant for improving human reasoning skills. Int J Hum Comput Stud. 2018. https://doi.org/10.1016/j.ijhcs.2018.02.005.
32. Coronado M, Iglesias AC, Carrera Á, Mardomingo A. A cognitive assistant for learning java featuring social dialogue. Int J Hum Comput Stud. 2018. https://doi.org/10.1016/j.ijhcs.2018.02.004.
33. Garcia-Serrano MA, Martinez P, Hernandez ZJ. Using AI techniques to support advanced interaction capabilities in a virtual assistant for e-commerce. Expert Syst Appl. 2004;26:413–26.
34. Ackerman MS, Dachtera J, Pipek V, Wulf V. Sharing knowledge and expertise: the CSCW view of knowledge management. Comput Support Coop Work CSCW Int J. 2013. https://doi.org/10.1007/s10606-013-9192-8.

35. Weidong F, Xi TM, Frazer JH. Constructing an intelligent collaborative design environment with distributed agents, 8th Int. Conf. Comput. Support. Coop. Work Des. (n.d.).
36. Wu S, Ghenniwa H, Zhang Y, Shen W. Personal assistant agents for collaborative design environments. Comput Ind. 2006. https://doi.org/10.1016/j.compind.2006.04.010.
37. Zhang Y, Ghenniwa H, Shen W. Agent-based personal assistance in collaborative design environments. Int. Conf. Comput. Support. Coop. Work Des. (n.d.).
38. Ferguson G, Quinn J, Horwitz C, Swift M, Allen J, Galescu L. Towards a personal health management assistant. J Biomed Inform. 2010;43:S13–6. https://doi.org/10.1016/j.jbi.2010.05.014.
39. Reumann M, Ieee SM, Giovannini A, Nadworny B, Auer C, Girardi I, Marchiori C. Cognitive DDx assistant in rare diseases. 2018:3244–7.
40. Rincon J, Costa A, Novais P, Julian V, Carrascosa C. A new emotional robot assistant that facilitates human interaction and persuasion. Knowl Inf Syst. 2018. https://doi.org/10.1007/s10115-018-1231-9.
41. Carmien S, DePaula R, Gorman A, Kintsch A. Increasing workplace independence for people with cognitive disabilities by leveraging distributed cognition among caregivers and clients. In: Proc. Int. ACM Siggr. Conf. Support. Gr. Work, 2003. https://doi.org/10.1145/958175.958176.
42. Thompson P, Iqbal R, James A. Supporting collaborative virtual meetings using multi-agent systems. In: 2009 13th Int. Conf. Comput. Support. Coop. Work Des. (n.d.).
43. Thompson P, James A, Iqbal R. Agent based facilitator assistant for virtual meetings. In: Proc. 2011 15th Int. Conf. Comput. Support. Coop. Work Des. CSCWD 2011, 2011. https://doi.org/10.1109/CSCWD.2011.5960095.
44. Ogata H, Yano Y, Furugori N, Jin Q. Computer supported social networking for augmenting cooperation. Comput Support Coop Work. 2001. https://doi.org/10.1023/A:1011216431296.
45. Aday LA. Health status of vulnerable populations. Annu Rev Public Health. 1994;15:487–509. https://doi.org/10.1146/annurev.pu.15.050194.002415.
46. Ahmetovic D, Gleason C, Ruan C, Kitani K, Takagi H, Asakawa C. NavCog: a navigational cognitive assistant for the blind. In: Proc. 18th Int. Conf. Human-Computer Interact. with Mob. Devices Serv. - MobileHCI '16, 2016. https://doi.org/10.1145/2935334.2935361.
47. Wolters KM, Kelly F, Kilgour J. Designing a spoken dialogue interface to an intelligent cognitive assistant for people with dementia. Health Inform J. 2016;22:854–66. https://doi.org/10.1177/1460458215593329.
48. Saad U, Afzal U, El-Issawi A, Eid M. A model to measure QoE for virtual personal assistant. Multimed Tools Appl. 2016. https://doi.org/10.1007/s11042-016-3650-5.
49. Tokadlı G, Dorneich CM. Development of design requirements for a cognitive assistant in space missions beyond low earth orbit. J Cogn Eng Decis Mak. 2018. https://doi.org/10.1177/1555343417733159.
50. Gopher D, Braune R. On the psychophysics of workload: why bother with subjective measures? Hum Factors. 1984;26:519–32. https://doi.org/10.1177/001872088402600504.
51. Wickens CD. Multiple resources and mental workload christopher. Hum Factors. 2008;50:449–55. https://doi.org/10.1518/001872008X288394.
52. Reid GB, Nygren TE. The subjective workload assessment technique: a scaling procedure for measuring mental workload. Adv Psychol. 1988. https://doi.org/10.1016/S0166-4115(08)62387-0.
53. Wierwille WW, Casali JG. A validated rating scale for global mental workload measurement applications. Proc Hum Factors Soc. 1983;27:129–33. https://doi.org/10.1177/154193128302700203.
54. Vidulich MA, Tsang PS. Absolute magnitude estimation and relative judgement approaches to subjective workload assessment. In: Proc. Hum. Factors Soc. Annu. Meet. 1987. p. 1057–61.
55. Tsang PS, Velazquez VL. Diagnosticity and multidimensional subjective workload ratings. Ergonomics. 1996. https://doi.org/10.1080/00140139608964470.
56. Zijlstra FRH. Efficiency in work behaviour: a design approach for modern tools, Delft Univ. Press. 1993. ISBN: 90-6275-918-1.
57. Ma J, Jaradat R, Ashour O, Hamilton M, Jones P, Dayarathna VL. Efficacy investigation of virtual reality teaching module in manufacturing system design course. J Mech Des. 2018;141:012002. https://doi.org/10.1115/1.4041428.
58. Bernstein WZ, Ramanujan D, Kulkarni DM, Tew J, Elmqvist N, Zhao F, Ramani K. Mutually coordinated visualization of product and supply chain metadata for sustainable design. J Mech Des Trans ASME. 2015. https://doi.org/10.1115/1.4031293.
59. Starkey EM, McKay AS, Hunter ST, Miller SR. Piecing together product dissection: how dissection conditions impact student conceptual understanding and cognitive load. J Mech Des Trans ASME. 2018. https://doi.org/10.1115/1.4039384.
60. Yang C-H, Hwang S-L, Wang J-L. The design and evaluation of an auditory navigation system for blind and visually impaired. In: Proc. 2014 IEEE 18th Int. Conf. Comput. Support. Coop. Work Des. (n.d.).
61. Alharthi SA, Raptis GE, Katsini C, Dolgov I, Nacke LE, Toups ZO. Toward understanding the effects of cognitive styles on collaboration in multiplayer games. In: Proc. ACM Conf. Comput. Support. Coop. Work. CSCW, 2018. https://doi.org/10.1145/3272973.3274047.
62. Yamashita N, Kaji K, Kuzuoka H, Hirata K. Improving visibility of remote gestures in distributed tabletop collaboration. In: Proc. ACM Conf. Comput. Support. Coop. Work. CSCW, 2011. https://doi.org/10.1145/1958824.1958839.
63. Kalnikaite V, Ehlen P, Whittaker S. Markup as you talk: establishing effective memory cues while still contributing to a meeting. In: Proc. ACM Conf. Comput. Support. Coop. Work. CSCW, 2012. https://doi.org/10.1145/2145204.2145260.
64. Olsson S, Burns PCC. Measuring driver visual distraction with a peripheral detection task. NHTSA Driv Distraction Internet Forum. 2000. https://doi.org/10.1097/JCP.0b013e3182a59409.
65. Martens MH, van Winsum W. Measuring distraction: the peripheral detection task, TNO Hum. Factors. 1996. p. 1–7.
66. Brookhuis KA, van Driel CJG, Hof T, van Arem B, Hoedemaeker M. Driving with a congestion assistant; mental workload and acceptance. Appl Ergon. 2009;40:1019–25. https://doi.org/10.1016/j.apergo.2008.06.010.
67. Svensson E, Angelborg-Thanderez M, Sjöberg L, Olsson S. Information complexity-mental workload and performance in combat aircraft. Ergonomics. 1997. https://doi.org/10.1080/001401397188206.
68. Lemoine MP, Debernard S, Crevits I, Millot P. Cooperation between humans and machines: first results of an experiment with a multi-level cooperative organisation in air traffic control. Comput Support Coop Work. 1996. https://doi.org/10.1007/BF00133661.
69. Oreg S, Martin MM, Rubin RB, Heaton RK, Chelune GJ, Talley JL, Kay GG, Curtiss G. Wisconsin card sorting test manual: revised and expanded. Psychol Rep. 1993. https://doi.org/10.2466/pr0.1995.76.2.623.
70. Rhodes MG. Age-related differences in performance on the Wisconsin card sorting test: a meta-analytic review. Psychol Aging. 2004. https://doi.org/10.1037/0882-7974.19.3.482.
71. Fristoe NM, Salthouse TA, Woodard JL. Examination of age-related deficits on the Wisconsin Card Sorting Test.

Neuropsychology. 1997. https://doi.org/10.1037/0894-4105.11.3.428.
72. Ozonoff S. Reliability and validity of the Wisconsin card sorting test in studies of autism. Neuropsychology. 1995. https://doi.org/10.1037/0894-4105.9.4.491.
73. Channon S. Executive dysfunction in depression: the Wisconsin Card Sorting Test. J Affect Disord. 1996. https://doi.org/10.1016/0165-0327(96)00027-4.
74. Gold JM, Carpenter C, Randolph C, Goldberg TE, Weinberger DR. Auditory working memory and Wisconsin card sorting test performance in schizophrenia. Arch Gen Psychiatry. 1997. https://doi.org/10.1001/archpsyc.1997.01830140071013.
75. Lyvers MF, Maltzman I. Selective effects of alcohol on Wisconsin card sorting test performance. Br J Addict. 1991. https://doi.org/10.1111/j.1360-0443.1991.tb03417.x.
76. Naderifar M, Goli H, Ghaljaie F. Snowball sampling: a purposeful method of sampling in qualitative research. 2017. p. 1–6. https://doi.org/10.1510/icvts.2010.244582.
77. Puente AE. Wisconsin Card Sorting Test, Test Crit. 1985. p. 677–82.
78. Nevin JA. Signal detection theory and operant behavior. J Exp Anal Behav. 1969. https://doi.org/10.1007/s00221-011-2557-7.
79. Jahn G, Oehme A, Krems JF, Gelau C. Peripheral detection as a workload measure in driving: effects of traffic complexity and route guidance system use in a driving study. Transp Res Part F Traffic Psychol Behav. 2005;8:255–75. https://doi.org/10.1016/j.trf.2005.04.009.
80. Verwey WB. On-line driver workload estimation. Effects of road situation and age on secondary task measures. Ergonomics. 2000;43:187–209. https://doi.org/10.1080/001401300184558.
81. Patten CJD, Kircher A, Östlund J, Nilsson L. Using mobile telephones: cognitive workload and attention resource allocation. Accid Anal Prev. 2004;36:341–50. https://doi.org/10.1016/S0001-4575(03)00014-9.
82. Grant RC, Carswell CM, Lio CH, Seales WB. Measuring surgeons' mental workload with a time-based secondary task. Ergon Des. 2013;21:7–11. https://doi.org/10.1177/1064804612466068.
83. Stoet G. PsyToolkit: a software package for programming psychological experiments using Linux. Behav Res Methods. 2010. https://doi.org/10.3758/BRM.42.4.1096.
84. Stoet G. PsyToolkit: a novel web-based method for running online questionnaires and reaction-time experiments. Teach Psychol. 2017. https://doi.org/10.1177/0098628316677643.
85. Maulsby D, Greenberg S, Mander R. Prototyping an intelligent agent through Wizard of Oz. In: Proc. SIGCHI Conf. Hum. Factors Comput. Syst. - CHI '93. 1993. p. 277–84. https://doi.org/10.1145/169059.169215.
86. Riek L. Wizard of Oz studies in HRI: a systematic review and new reporting guidelines. J Hum Robot Interact. 2012;1:119–36. https://doi.org/10.5898/JHRI.1.1.Riek.
87. Klemmer SR, Sinha AK, Chen J, Landay JA, Aboobaker N, Wang A. SUEDE: a wizard of Oz prototyping tool for speech user interfaces, UIST (User Interface Softw. Technol. Proc. ACM Symp. 2000. p. 1–10.
88. Jou W, Beaulieu SM, Lim AK, MacDonald EF. A wizard-of-oz experiment to demonstrate water reduction and user training with an "autonomous" faucet. In: Int. Des. Eng. Tech. Conf. Comput. Inf. Eng. Conf. Proc. 2019. 2019.
89. Mitchell MS, Yu MC, Whiteside TL. Editorial: The tyranny of statistics in medicine: a critique of unthinking adherence to an arbitrary p value. Cancer Immunol Immunother. 2010. https://doi.org/10.1007/s00262-010-0859-4.
90. Vaske JJ, Gliner JA, Morgan GA. Communicating judgments about practical significance: effect size, confidence intervals and odds ratios. Hum Dimens Wildl. 2002. https://doi.org/10.1080/10871200214752.
91. Aiken KD. explorations in interpersonal trust development: the trust curve, 1999.
92. Lee DJ, See AK. Trust in automation: designing for appropriate reliance. Hum Factors. 2001;46:50–80. https://doi.org/10.1518/hfes.46.1.50_30392.
93. Aghaei B**.** Adaptive affective computing: countering user frustration. 2013.
94. Cohn AM, Hunter-Reel D, Hagman BT, Mitchell J. Promoting behavior change from alcohol use through mobile technology: the future of ecological momentary assessment. Alcohol Clin Exp Res. 2011. https://doi.org/10.1111/j.1530-0277.2011.01571.x.
95. Wilemon DL, Thamhain HJ. Team building in project management. In: Proc. Proj. Manag. Inst. Annu. Semin. Symp. 1979. p. 373–80.
96. Tolmie P, Crabtree A, Rodden T, Benford S. "Are you watching this film or what?": Interruption and the juggling of cohorts. In: Proc. ACM Conf. Comput. Support. Coop. Work. CSCW, 2008. https://doi.org/10.1145/1460563.1460605.