AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Neuroscience, Psychology, and Economics

ISSN: 1937-321X

2024, Vol. 17, No. 2, 119–144
https://doi.org/10.1037/npe0000191

Consumer Decisions in Virtual Commerce: Predict Good Help-Timing Based on Cognitive Load

Tobias Weiß[1] and Jella Pfeiffer[2]

[1] Fachbereich 02 Wirtschaftswissenschaften, Justus Liebig University Gießen

[2] Abteilung VII, Wirtschaftsinformatik I, University Stuttgart

The retail sector is steadily moving toward virtual commerce (v-commerce), and the process has recently gained momentum. With the latest developments in headset technology and the rise of artificial intelligence, virtual shopping has become relevant for an increasing number of products. In this article, we present a study that combines consumer behavior research, eye tracking, electrocardiography, machine learning, and the application of virtual reality. Fifty participants were invited to experience a virtual scenario, perform multiple mentally demanding tasks, and make a purchase decision for a product from one of two different product categories. In a post hoc video analysis based on the first-person view, participants determined different points in time when they would have appreciated help from an algorithmic user assistance system or a digital human agent. Our statistical analysis suggests that the desired help-timing depends on the product category. For fast-moving consumer goods, algorithmic help was requested particularly early. Furthermore, we collected eye-tracking and electrocardiographic data to build and evaluate a predictive classification model that differentiates between three levels of cognitive load. The trained machine learning algorithm aims to classify cognitive load during decision making, which may indicate the right time to offer help. Our findings provide evidence that eye movements, in particular, allow service providers to determine a good moment to approach consumers during their shopping experience.

Keywords: consumer behavior, eye tracking, virtual reality, electrocardiography, machine learning

Supplemental materials: https://doi.org/10.1037/npe0000191.supp

The popularity of online shopping has transformed traditional brick-and-mortar stores into highly competitive virtual marketplaces (Bourlakis et al., 2009). While technological advances provide new opportunities for consumers to visualize and experience their environment, new business rules pose challenges for retailers seeking to provide engaging and meaningful experiences (Reinartz et al., 2019). With the proliferation of immersive technologies such as virtual reality (VR), the idea of the Metaverse continues to fascinate many people. For immersive shopping scenarios, knowledge about cognitive processes can help to design highly personalized user assistance systems (UAS).

Tobias Weiß https://orcid.org/0009-0007-1417-8044

The work was funded by the Deutsche Forschungsgemeinschaft (German Research Foundation)—GRK2739/1—Project Nr. 447089431 (Tobias Weiß)—Research Training Group: KD²School—Designing Adaptive Systems for Economic Decisions. The authors encourage readers to reproduce and expand upon their results and analysis. The authors released their complete pseudoanonymized data set and source code (Weiß, 2023).

Correspondence concerning this article should be addressed to Tobias Weiß, Fachbereich 02 Wirtschaftswissenschaften, Justus Liebig University Gießen, Licher Straße 74, 35394 Gießen, Germany. Email: tobias.weiss@wirtschaft.uni-giessen.de

Decision support systems are an elemental tool for retailers that can severely impact their business success (Shim et al., 2002). As a subclass, UAS can be seen as a joint element which "bridge[s] the gap between the system's functionalities and the human's individual capabilities with the goal of positively influencing task outcomes" (Morana et al., 2020, p. 189).

Due to the need for enhanced consumer experiences, several studies suggest that the provision of personalized user assistance will become highly relevant in v-commerce scenarios (B. Chen & Yang, 2022; Guo & Elgendi, 2013; Z. Zhang et al., 2013). UAS in e- and v-commerce include conversational agents (Heßler et al., 2022), recommendation systems (Xiao & Benbasat, 2007), and virtual assistants (Raut et al., 2023). In general, user assistance leverages analytics, data, and technology to help consumers make informed decisions about various aspects of their purchases. Examples of algorithmic help offerings include displaying the most relevant product reviews from other consumers (Pan & Zhang, 2011) or assisting with interactive decision aids (Häubl & Trifts, 2000; Pfeiffer, 2011).

With the ability to collect data on neurophysiological responses in VR, new opportunities arise to create intelligent UAS that adapt to the individual's state. Machine learning (ML) plays a crucial role when building these new UAS as it provides the basis for artificial intelligence (AI) steering the system. An intelligent, ML-based, adaptive system can learn about consumer search motives (Pfeiffer et al., 2020) using eye tracking (ET). Among the latest VR headsets, the most common biosensors are ET cameras. For this reason, we utilize ET as the main neurophysiological sensor to detect visual attention and predict cognitive load. However, recent research-grade VR headsets offer further data sources, like electrocardiographic (ECG) sensors, and we forecast that a variety of different sensors will be available, as well as additional wearables. For example, electroencephalography (EEG) earbuds (Athavipach et al., 2019), for which a major tech company recently patented a design.

One aspect that might help to create a good, highly personalized user experience (and therefore impact the success of these sales interactions) is the time when consumer assistance is invoked (Friemel et al., 2018). Adequate timing can influence consumers' attention (Bailey & Konstan, 2006), perceived relevance, trust, and urgency, and could be an enabler for UAS providers to beat the competition (Meurisch et al., 2020). Peukert et al. (2020) outlined how important it is to display a UAS with good timing. They proposed a decision-phase-based detection algorithm and compared it with previously suggested decision-phase models (Gidlöf et al., 2013; Russo & Leclerc, 1994). However, they used simple gaze pattern rules to determine the phases, such as the first refixation on a product. A good timing to approach a consumer, however, depends on several factors, including their mental state (e.g., in the form of cognitive overload, personality, and habitual purchasing patterns). By carefully timing interactions, we claim that both consumers and providers can benefit due to the avoidance of dissonance between intended help offering and, in the worst case, perceived annoyance. While further previous work focused on assistance timing in generic software interface tasks, like finding appropriate software functionalities to alter an image (Ginon et al., 2016), this study is particularly geared towards the consumer decision-making context in v-commerce.

In this article, we investigate cognitive load and its capability as an indicator to determine a good timing to engage with consumers in a shopping scenario. Previous work has identified cognitive load as a key mental state for decision making (Deck & Jahedi, 2015). In line with findings from the educational domain (Vaessen et al., 2014), we hypothesize that high levels of cognitive load can make it more difficult for consumers to understand and solve decision problems on their own, leading them to seek help (in the form of an algorithmic support system or a digital human agent, i.e., a human sales representative controlling an avatar in the virtual shopping environment). Low levels of cognitive load might increase consumers' confidence and ability to solve problems independently, reducing the likelihood that they seek or appreciate help but rather want to browse the store independently. We argue that by estimating the cognitive load level during a consumer's purchase decision, it might be possible to determine a good timing to start an interaction. To account for varying levels of product knowledge, we employ two distinct products from two different categories: a fast-moving consumer good and a technology product. We expect differences between the product categories regarding desired help-timing. Thus, the research questions read as follows:

1. When is the desired help-timing for algorithmic user assistance compared to

the desired help-timing for a digital human agent in different shopping scenarios?
2. How does product knowledge influence the desired help-timing?
3. Is desired help-timing related to cognitive load and if yes, how can cognitive load be used to determine a good intervention timing?

We investigate these questions in an experimental VR environment, which gives our study particular relevance in the light of the latest developments in the retail domain towards v-commerce. VR can improve consumer experiences (Moghaddasi et al., 2021) and offer high external validity while maintaining experimental control (Meißner et al., 2019). Furthermore, the used high-end VR headset allows us to collect gaze patterns and pupillometry in an unobtrusive and precise way. To answer our questions, we draw from two data sources. Both ET and ECG serve as indicators of cognitive load (Haapalainen et al., 2010). This article mainly builds upon two works. First, Peukert et al. (2020) have used ET to distinguish decision phases by using simple gaze patterns. These phases might indicate a good point in time when users seek help, but a connection between decision phases and help-timing was not investigated in their article. Second, Pfeiffer et al. (2020) have estimated search motives based on fixations and their statistical moments. To complement the fixation data, we additionally include blinks, saccades, and pupillometry in the feature set. Additionally, we use ECG as secondary neurophysiological sensor. We extend this existing stream of literature on consumer behavior in VR by focusing on the desired support type and good intervention timing.

Our contributions are twofold. First, we show that desired help-timing depends on whether the help is provided by an algorithmic UAS or a digital human agent. The desired help-timing also depends on the product category being purchased. As a result, when designing good shopping assistance, companies should be aware of this heterogeneity and strive for a high degree of personalization and context awareness of the shopping situation. Second, we investigate cognitive load as an indicator to estimate the timing of assistance by using ET and ECG. The study demonstrates how ET and ECG can be used as features for shallow and interpretable ML models to predict optimal assistance offers. Overall, this article emphasizes the transformative nature of v-commerce and the high relevance of leveraging the recently available extended set of biosensors. We provide valuable practical guidance on how to approach the v-commerce transition and take advantage of the technological opportunities.

Related Work

Cognitive Load

The mental effort or capacity required to process and understand information is referred to as Cognitive Load (CL). Originating in psychology and education, Cognitive Load Theory (CLT) explains how the human brain processes information during learning and problem-solving (Plass et al., 2010; Sweller, 2011). CLT suggests that humans have a limited amount of mental capacity (Miller, 1956) and that the difficulty of a task can affect how much of this capacity it occupies. Furthermore, CLT can be applied to decision making when choosing among several options (Deck & Jahedi, 2015). A variety of biosensors and ML techniques are available for measuring CL (Seitz & Maedche, 2022). To minimize the negative impact of CL on decision making, it is a viable option to simplify decision-making processes and reduce the amount of information that must be processed at a time (Todd & Benbasat, 1994). Today's software solutions can reduce CL and improve decision making by providing help from a virtual agent (Brachten et al., 2020). Another option is breaking down overwhelming decision-making tasks into smaller, more manageable parts. Still, task optimization and atomization are no panacea. Even if the amount of options is limited, empirical results suggest that high CL levels can negatively impact the quality of decision making (Allen et al., 2014; Dewitte et al., 2005). These studies consistently showed how a high CL level can lead to an increased likelihood of making errors in different task arrangements. Given this critical relation between CL and increased error rates, it is not surprising that marketing and shopping contexts are important domains to apply CLT (Grzyb et al., 2018; Schmutz et al., 2010; Wang et al., 2014). For example, a CLT-informed UAS can improve consumers' abilities to understand and process information about a product or service they are considering buying. By reducing CL, product vendors can foster a positive shopping experience

for their consumers. Building on the CLT principles, shop providers can actively design a UAS that increases their consumers' motivation and ability to seek help when needed. By making it easier for consumers to seek help when needed or even offering the required help with perfect timing, companies can improve consumer satisfaction and reduce costs associated with providing assistance (Caruelle et al., 2023). Overall, CLT can provide a basis for understanding how different levels of CL influence consumers' motivation and ability to seek help. We hypothesize that after an initial exploration/orientation phase, consumers want to mitigate the imposed CL burden and value support. We further believe that CL can help to identify the moment when consumers engage with the product, viewing and comparing attributes or details. Such behavior indicates an increased likelihood of open questions. These questions could be answered by an algorithmic support system or a digital human agent.

Eye Tracking

Gaze patterns are suitable for tracking visual attention (Duchowski, 2017), but their analysis relies on the eye–mind hypothesis by Just and Carpenter (1980), which assumes that human cognitive processes can be observed by their associated gaze patterns. However, it is evident that individuals can deliberately look at a certain position while thinking about something else (Anderson et al., 2004). Nonetheless, experimental findings indicate the validity of the eye–mind hypothesis in numerous scenarios (Holmqvist et al., 2011). Important movement-related gaze metrics are fixations and saccades. A fixation is a stationary state of the eyes and can last from milliseconds to seconds, while saccades are rapid eye movements between fixations.

Pupillometry investigates the changes in pupil dilation and frequently serves as an estimator for CL (Hess & Polt, 1964; Holmqvist et al., 2011; Kahneman, 1973). In natural environments, pupillometry is not reliable for determining CL because small deviations in the lighting conditions have a strong impact on pupil dilation (Laeng et al., 2012). In a virtual environment experienced by an individual using a VR headset, lighting confounds can be mitigated because the closed head-mounted display (HMD) cover offers fully controllable scene lighting.

Electrocardiography

ECG records the electrical activity of the heart, which emits a group of waves called PQRST (Goy, 2013). Research has applied ECG to investigate various aspects of consumer behavior and is commonly used in combination with other biometric tools (Harris et al., 2018). Human–computer interaction research assesses additional factors, such as the usability of user interface design (Lee & Seo, 2010) and emotional engagement with presented information (Ferdinando et al., 2016). ECG can serve as an indicator for CL (Haapalainen et al., 2010; Hughes et al., 2019). Data collection is typically performed with high frequency using electrodes that are attached to the skin.

Virtual Reality

In VR, the real-world environment is replaced as comprehensively as possible. The main goal of VR is to create realistic but completely virtual experiences with a high level of (tele-)presence for the users (Cummings & Bailenson, 2016). An early HMD, as it is common today, was already developed by Sutherland (1965). Another option to create virtual spaces is a *CAVE automatic virtual environment* (a recursive acronym), a cube-shaped room with projections on its walls (Cruz-Neira et al., 1992). Today, HMDs are common, and some models can even show mixed reality, which means everything on a spectrum from slightly augmented to fully immersive experiences. It is possible to combine an HMD with a variety of different sensors and cameras, particularly ET (Pfeiffer et al., 2020), which leads to many interesting research opportunities. Moreover, VR mitigates the trade-off between experimental control and ecological validity (Meißner et al., 2019).

VR has changed the landscape of v-commerce, ushering in a new era of immersive and personalized shopping experiences (Evans & Wurster, 1999). The technology might transform the way consumers interact with products and purchase them online by providing a more engaging and lifelike representation. VR showrooms allow consumers to view products in three dimensions, enabling a more informed decision-making process. In addition, VR has enhanced the social aspect

of v-commerce through shared virtual spaces where friends or family can shop together and share opinions in real time (H. Zhang et al., 2014). A recent review by Branca et al. (2023) provides a comprehensive overview of different literature streams that address v-commerce. The authors identify four key research streams: consumers, products, product testing, and VR compared to other conditions. As our study mainly focuses on desired help-timing, it fits into the consumer category. However, we propose to introduce a fifth label called *sales agents*, which covers related research. We argue that the interface between provider and consumer is a key success factor that needs increased attention. Table 1 provides a list of selected previous consumer behavior experiments in VR. It briefly describes the experimental setups and contributions, and allows the reader to understand the contribution and positioning of our article.

Method

Experimental Design

The experimental setup was based on a virtual showroom in VR. Participants performed generic CL tasks of three difficulty levels and a subsequent purchase decision. The experiment focused on the utilitarian aspect of consumer behavior, as we asked participants to make decisions based on a set of criteria, leaving little room for their own hedonic motives. A web-based questionnaire on a desktop computer complemented the VR recordings. For all experiment sessions, we collected ET and ECG data.

To answer the first research question, we examined consumers' desired help-timing for an algorithmic UAS versus a digital human agent. To identify potential differences across product categories, we used two product sets of four items. One set represented technology products (3D printers), and the other set represented fast-moving consumer goods (washing powders). We asked participants to identify good intervention timings for the two different types of help providers, an algorithmic UAS and a digital human agent because participants might perceive relevant differences between these help providers. We argue that an algorithmic UAS may appear earlier during a decision-making process than a digital human agent because it is comparatively inexpensive. For the intervention of a digital human agent, timing is critical because it translates into substantial costs for human resources on the seller's side. Providers should, therefore, be confident that an engagement is desired and that it takes place at the appropriate time.

As an exploratory aspect related to the first research question, we also wanted to identify the specific desired help type for algorithmic user assistance. In other words, do users prefer interactive decision aids, recommendations, or other algorithmic help types? This insight may guide practitioners in deciding which system type to implement in a certain scenario.

To investigate the second research question, we compared participants' product knowledge for the different product categories and examined its relationship with desired help times. We expected low product knowledge for the 3D printers because they are niche products, whereas a broad range of participants should be familiar with different washing powders. However, it was not clear what effect this (un-)familiarity would have on desired help times.

To control for possible confounding, we collected the participants' demographic information, personality traits, and their general attitude toward sales representatives. We also asked the participants about their product involvement but expected little difference because the monetary incentive for solving the purchase task was the same in both the washing powder and 3D printer scenarios.

To answer our third research question, which aims to increase the understanding of CL in relation to the point in time when consumers want help, we measured CL levels that participants experienced when solving three generic tasks of low, middle, and high complexity before transitioning to the actual purchase task. To verify the difficulty levels, we controlled for subjectively perceived complexity during the generic tasks. Using the recorded ET and ECG data, we trained an XGBoost model to predict the CL level during a short period prior to the desired help-timing.

All virtual scenes were implemented using the Unity 2021.3 game engine. Participants experienced our virtual environments using a Varjo VR 3 HMD with Valve Index controllers. This headset offered high-frequency ET capability with a sampling rate of up to 200 Hz, and its display resolution of $2{,}880 \times 2{,}720$ pixels per eye led to high visual immersion. The ET sensor was

Table 1
Related VR Experiment Categorization

Study	Setup	Contribution
Bigné et al. (2016)	$N = 41$ CAVE ET data Spatial data Questionnaire	This study investigates brand preferences for fast food products and suggests that high attention to a brand and slow eye movements between brands lead to additional brand purchases. The applied method consists of regressions with aggregated parameters related to the entire decision-making process.
Martinez-Navarro et al. (2019)	$N = 178$ HMD Questionnaire	The authors compare the effectiveness of different VR formats and devices. They find that virtual stores are more effective in generating cognitive and conative responses. They apply a structural equation model that suggests a dual path via brand recall and presence, both of which influence consumers' purchase intention in virtual stores.
Meißner et al. (2020)	$N = 132$ HMD Questionnaire	This article compares high-immersive (using an HMD) and low-immersive shopping environments (using a desktop computer) and examines consumers' variety-seeking, price sensitivity, and choice satisfaction. In an incentive-aligned choice experiment, participants make repeated purchase decisions for cereal products. The statistical analysis suggests that consumers are less price sensitive and seek more variety in highly immersive environments.
Pfeiffer et al. (2020)	$N = 50$ CAVE ET data Questionnaire	The authors investigate two classic shopping motives: goal-directed search and exploratory browsing. They compare decisions in a real-world supermarket with decisions in a virtual reality supermarket. They collect ET data on which they train three shallow ML models. They found that an ensemble method can classify the two motives with about 90% accuracy.
Alzayat and Lee (2021)	$N = \{48, 35\}$ HMD Questionnaire	Using two VR stages and an Amazon mturk questionnaire, the authors investigate the differences in hedonic purchase value between a VR retail environment and a website. Their analysis comprises three structural equation models. The results suggest that a VR retail environment is more appropriate for products that are perceived as an extension of the body (e.g., tools) rather than a representation of the body (e.g., clothing).
Huang et al. (2021)	$N = 80$ HMD Brain activity Questionnaire	This article focuses on search behavior, which is involved in the evaluation phase of each decision-making process. The authors investigate the congruence or incongruence between text and color of flavor labels on product packaging. They provide evidence for a color–flavor incongruence effect in visual search and correlate it to the violation of user expectations. The method involves subsequent VR and functional magnetic resonance imaging phases, which the authors analyze using multiple regressions and regional homogeneity analyses, respectively.
H. Park and Kim (2023)	$N = 196$ HMD Questionnaire	This research examines how offering a virtual try-on in Augmented Reality, a 3D store on a desktop computer, and a VR store affect consumers' purchase intentions. The study also analyzes how thinking more deeply about an item influences the decision-making process in different shopping scenarios (searching vs. browsing). Results indicate that purchase intentions are highest when participants browse in the VR condition. A moderated mediation analysis supports the hypothesis that cognitive elaboration mediates purchase intentions for those consumers in the browsing mode, while such a mediating effect was absent in the searching mode.
Schnack et al. (2021)	$N = 36$ HMD EEG data Spatial data Purchase data Questionnaire	This study compares instant teleportation with motion-tracked walking in VR and investigates whether different locomotion techniques correlate with altered shopping behavior. Using a split-sample experimental design, the authors apply electroencephalography (EEG) to track emotional states such as stress. In the scenario, participants experience a VR grocery store. Overall, the results suggest that different locomotion techniques have no impact on the consumers' emotional state and engagement. However, different spatial movement patterns are noticeable when comparing the different conditions.

(*table continues*)

Table 1 (*continued*)

Study	Setup	Contribution
Harz et al. (2022)	$N = 210$ HMD Questionnaire	The authors report on a combination of a real-world field study, which is followed by a laboratory experiment. They examine how durable goods companies can use VR for new product development and how VR can improve prelaunch sales forecasting. One of the three experimental conditions takes place in VR; the other conditions take place online and in the real world. The analysis of variances suggests that sales forecasting in VR provides the most accurate predictions compared to the other conditions. Moreover, it confirms the first evidence of the field study that VR correlates with a more consistent consumer behavior and that virtual reality might create superior behavioral consistency compared to the real world.
Our work	$N = 50$ HMD ET data ECG data Questionnaire	In contrast to previous work, our study focuses on the desired help-timing in a VR scenario for an algorithmic UAS versus a digital human agent. As a second dimension, we compare technical products (3D printers) with fast-moving consumer goods (washing powders). We present the statistical analysis of our questionnaire and apply a machine learning approach to identify a good intervention timing. During our experiment, participants solve CL-inducing tasks before making a purchase decision. ET and ECG provide the features for an ML classifier. Algorithmic help was requested particularly early for the washing powder. The results further indicate that CL-based classification works for the desired help-timing of an algorithmic UAS but not for a digital human agent. The approach could be refined to invoke an AI agent based on a fine-tuned large language model, that has in-depth product knowledge.

Note. VR = virtual reality; ET = eye tracking; HMD = head-mounted display; CAVE = CAVE automatic virtual environment; ML = machine learning; CL = cognitive load; UAS = user assistance systems; AI = artificial intelligence.

calibrated at the beginning of each experiment stage using a five-dot calibration protocol. For ECG recording, a wireless bioPLUX device captured signals throughout the experiment with a sampling rate of 1,000 Hz. To be able to clarify possible confounds post hoc during data analysis, we additionally recorded all experimental sessions on video using a room camera. Overall, the experiment followed a between-subjects design (regarding the two product categories) and included several questionnaire parts that alternated with the VR stages. Mandatory VR breaks for the questionnaires had additionally reduced the risk of cybersickness (Davis et al., 2014) and exhaustion of the participants within the VR environment.

Participants

Our self-hosted online registration platform (Bock et al., 2014) helped to recruit participants and manage the experiment sessions. Additionally, we actively solicited participation from students on our campus. Participation requirements were an age between 18 and 65 years and good command of English and German. Furthermore, we only accepted participants with normal or corrected-to-normal vision. Participation compensation was €10 fixed plus a performance-based component of up to €5.5. After arriving at the lab, participants signed a consent form. It ensured the participants' basic knowledge of the experimental procedure, informed them that the experiment complied with ethical standards, and required them to grant permission to publish their pseudonymized data as an open-source data set.

Behavioral Measurements

We measured all questionnaire items on a 7-point Likert scale. In terms of demographics, we tracked participants' age, gender, and occupation. To estimate personality traits, we used the Big Five Index–10 short scale (Rammstedt et al., 2013), which allows for the evaluation of personality traits with acceptable validity in a compact manner. We measured the general desire to interact with a salesperson using eight items validated by Lee and Dubinsky (2017). To collect self-assessments about CL, for both the multitasking and decision stage, we asked participants to answer the six-item National Aeronautics and

Space Administration (NASA) Task Load Index (TLX) questionnaire (Hart, 2006; Hart & Staveland, 1988). Overall, four TLX batteries were collected per participant, one for each of the three generic CL task difficulty levels and one for the purchase decision. The product knowledge scale, consisting of three items, was adapted from C.-W. Park and Moon (2003) to fit the presented products (see online Supplemental Material). Moreover, the questions regarding participants' product involvement comprised 20 bipolar items (Zaichkowsky, 1985).

Neurophysiological Measurements

To generate features for the ML model from the collected sensor data, we aggregated the raw ET and ECG recordings. The extracted features are listed in the Supplemental Table 6 for ET features and in Supplemental Table 7 for ECG features.

For the ET data, we utilized both gaze-based metrics and pupillometry. Gaze events, namely fixations, saccades, and blinks, were created using a Velocity–Threshold Identification approach as described by Salvucci and Goldberg (2000). For saccades, we set 50°/s as the lower angular speed threshold (Holmqvist et al., 2011). We limited fixation durations to 0.1 s as the lower threshold and 10 s as the upper threshold (Duchowski, 2017). After creating the gaze events, we aggregated statistical moments to determine if attention was directed to different areas of interest (AOI, e.g., a product) and how often attention shifted between different AOIs. For pupillometry, we used the pupil–iris ratio of the dominant eye and complemented the gaze events with this information.

Using the raw ECG data, we extracted time- and frequency-domain-related features that covered different aspects of the heart rate and its variability (HRV) in linear and nonlinear representations (Chanel et al., 2019; Pham et al., 2021; Xiong et al., 2020). Regarding ECG feature selection, we rely on a recent review that covers the "most up-to-date and commonly used HRV indices" by Pham et al. (2021). Due to our relatively short task periods, some of the common HRV measures could not be investigated, such as the standard deviations of average heartbeat intervals, which compare longer segments (by default, 1, 2, and 5 min).

Overall, a crucial step for feature engineering was setting the time window size because it determined how the features were aggregated. For the ET-related features, we evaluated six different window sizes (3, 5, 7, 10, 15, 30 s, where 30 s is the full trial duration), which yielded equally long segments without overlapping or artificial padding.

Further assumptions are necessary for the ET postprocessing. An average fixation lasts about 0.3 s (Holmqvist et al., 2011) and average blinks and saccades are even shorter. Thus, we argue that 3 s yield enough data to calculate meaningful statistical moments in many cases. Considering increasing window sizes makes sense because CL might not be present from the onset of the task. Comparing different parts of a trial could yield a good contrast, such as the first versus the second half of a trial.

For ECG measurements, we only considered the full trial length (30-s windows). For shorter periods, only a limited set of features is computable, such as heart rate variability (HRV), while several features from the frequency domain and nonlinear domain suffer from numeric instabilities.

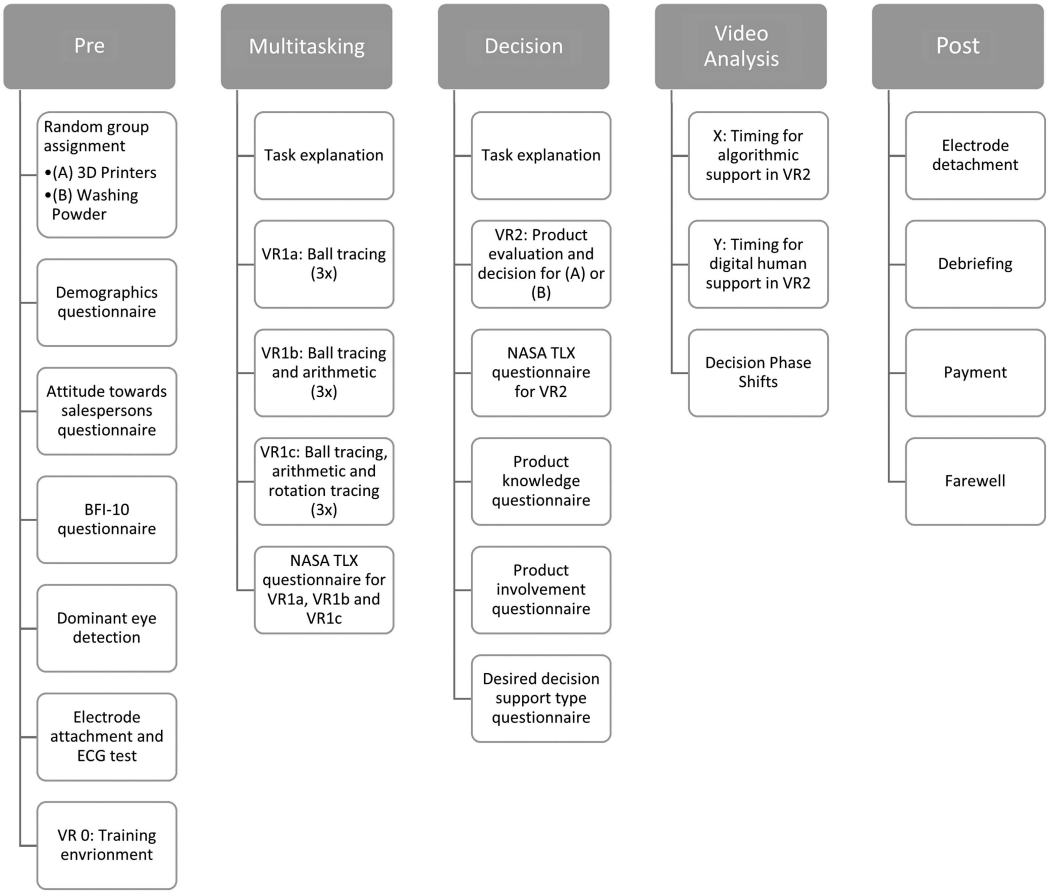
Procedure

The experiment lasted approximately 80 min, and it consisted of five different stages, as shown in Figure 1. The stages were streamlined with a web-based questionnaire on a desktop computer, which alternated with the VR scenes and guided participants through the different stages from start to end. During the prestage, our participants completed an onboarding procedure and answered general questions. A multitasking stage followed in which participants performed nine generic CL tasks (with three levels of complexity: easy, medium, and hard). A decision stage followed, in which participants made a product purchase to meet a list of given criteria. A video-analysis stage followed, during which participants retrospectively analyzed their first-person view during the purchase decision. A final poststage, in which participants went through our offboarding procedure, concluded the experiment.

Prestage

We randomly assigned arriving participants to one of two groups by flipping a coin and starting the corresponding questionnaire on the computer. In the subsequent decision stage, Group A was assigned to decide upon 3D printer products, and Group B was assigned to washing powder products. A welcome screen explained the general

Figure 1
Experiment Procedure



Note. ECG = electrocardiographic; NASA TLX = National Aeronautics and Space Administration Task Load Index; VR = virtual reality.

purpose and modalities of the experiment. Before continuing, we asked the participant to read and sign our consent form. Only after accepting the terms of the experiment, participants were asked to provide demographic data, information about their personality traits, and to answer questions about their general attitude towards salespersons. Next, we determined their dominant eye using the Miles test (Miles, 1929). For ECG data acquisition, we asked participants to go to the restroom to attach electrodes to their bodies according to a reference picture, and connect them to the transmitter. We decided to triangulate the heart in a wide triangle, spanning from the shoulders to the hip, to receive a high-quality signal that is robust to noise caused by body movements.

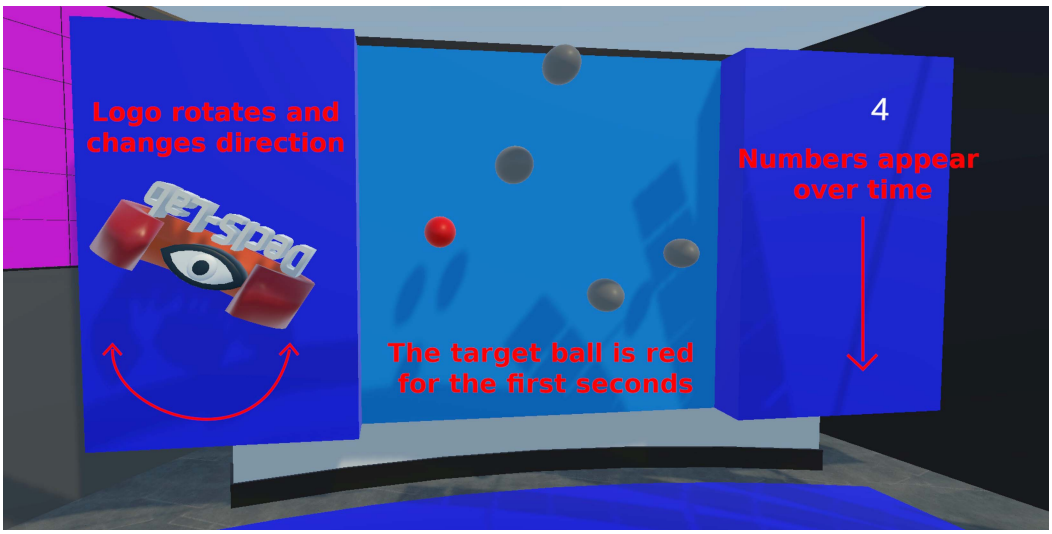
Next, we explained the VR hardware, controller usage, ET calibration procedure, and the upcoming task. Then, we familiarized participants with movement, teleportation, and interaction using a training environment very similar to the subsequent task environments. The training scene consisted of the same showroom, which was later used for the CL and decision environments. Participants were asked to use two in-world buttons that invoked the appearance of example models: one low-quality model with low polygon count and single-colored texture and one high-quality model with high polygon count and high-fidelity texture. Additionally, participants were asked to interact with a menu that started a timer and transitioned to the next stage after successful activation.

Multitasking Stage

To generate different generic CL levels, we designed a gamified CL task with three difficulty levels, as shown in Figure 2. This task was inspired by the work of Siegel et al. (2021). It consisted of three components—ball tracing, arithmetic, and rotation tracing. In the easy variant, participants had to trace one out of five moving balls. The target ball was colored red for 10 s. Afterward, the trial began, and the target ball changed its' color to the same gray as the other four balls. All five balls moved around pseudorandomly within a predefined area for 30 s. Finally, all balls stopped moving and displayed an identifying number. Participants then had to press a button labeled with the corresponding number to indicate which ball they considered as the target.

A text message informed the participants whether the answer was correct or not, and the task was reset after a short waiting time. The medium variant was more difficult as it included the easy variant but additionally introduced an arithmetic component. To the right of the ball tracing area, small pseudorandom numbers (ranging from −10 to 10) appeared sequentially on the wall within a pseudorandom time interval, and the participants had to aggregate them while still tracing the ball in parallel. At the end of each trial, a slider was presented with which the calculated sum could be entered. An additional text message informed the participants whether the answer was correct or not. The hard variant was even more difficult as it included the medium variant but additionally introduced a rotation tracking component. To the left of the ball tracing area, a spinning logo appeared, changing its rotational direction between clockwise and counterclockwise within pseudorandom time intervals. Participants had to count the amount of rotational direction changes in addition to the ball tracing and arithmetic components. After 30 s of trial time, participants saw an additional slider to enter the counted number of rotational direction changes. All difficulty levels were repeated three times, and we incentivized the correct completion of each trial by increasing participants' performance based extra payment by €0.5, if all components of a trial were answered correctly. Before these real trials started, all participants performed a training round in which they experienced the hard variant but without monetary incentive. During the training round, they could familiarize themselves with the task and ask questions. However, repetition was not possible. Then, they began with the easy variant, followed by the medium and hard variant, until all

Figure 2
Multitasking Virtual Reality Environment



Note. See the online article for the color version of this figure.

nine trials were completed. Afterward, participants took a VR break and continued the desktop-based questionnaire, which sequentially asked for their perceived task difficulty at all three levels.

Decision Stage

Both groups were presented with different task descriptions to create a realistic situation. Group A was asked to imagine being part of a board game designer team that needed a 3D printer to evaluate their game design. Group B was asked to imagine being a member of a residential community and being responsible for weekly grocery shopping, which included buying washing powder (see online Supplemental Material for the exact wording of both task descriptions). To incentivize the decisions and increase external validity, participants had the chance to gain one additional Euro performance-based participation compensation if their product choice matched a previously determined team decision. This team decision was negotiated by a group of five individuals in advance of the experimental sessions.

In the virtual environment, participants first saw a blackboard containing the requirements specified by their imaginary peers. We designed these requirements so that the difficulty level matched among the groups (see Supplemental Table 8). To this end, we chose three easy and three hard decision criteria. We considered attributes that were obvious by looking at the product packaging or the product description from a distance. On the other hand, we considered attributes as hard for which participants either had to interact with the product (e.g., by starting or turning it) or needed further information to be able to judge the product. An example of a required interaction is that the print quality of a 3D printer could only be determined by pressing the print button and looking at the printed object. An example of a criterion that needed more information is whether a washing powder is environmentally friendly. This is because the roommates could have been looking for environmentally friendly packaging, environmentally friendly product ingredients, or both. We believe that external help could be strongly appreciated to clarify the requirements for both groups.

To begin the decision phase after memorizing the requirements, participants had to press a start button that concealed the requirements on the blackboard and displayed the products on a table behind them (refer to Figure 3 for Group A and Figure 4 for Group B). After this, participants could approach and engage with the products. To make their decision, participants of Group A had

Figure 3
3D Printer Decision Virtual Reality Environment

Note. See the online article for the color version of this figure.

Figure 4
Washing Powder Decision Virtual Reality Environment

Note. See the online article for the color version of this figure.

to choose the respective 3D printer name from a drop-down menu and click a purchase button, while participants of Group B had to put the desired washing powder into a shopping cart next to the product table. After making a choice and detaching the HMD, participants continued to answer questions about their product knowledge, product involvement, task difficulty, and the preferred type of help for algorithmic user assistance (from a list of five common algorithmic user assistance types, see the online Supplemental Material).

Video-Analysis Stage

During this stage, participants answered time-related questions about their decision phase. Two questions regarding the desired timing for user assistance in the form of (X) an algorithmic UAS and (Y) a digital human agent (for exact wordings, see online Supplemental Material). These questions were displayed sequentially, and their order was randomized to avoid possible confounds induced by any static order. To find the corresponding timestamps, participants watched a video that showed their first-person view during the previous decision stage and also displayed a gaze dot indicating their visual attention. Participants then selected the most appropriate moment for the assistance to appear and entered the corresponding timestamp in the questionnaire.

Poststage

We asked participants to go to the restroom and detach the ECG transmitter and electrodes. Then, we continued with a debriefing (explanations about the experiment's purpose) and answered questions. Finally, we issued the participants' compensation and wished them farewell.

Results

The data analysis was performed in python 3.7 using neurokit2 0.2.3 (Makowski et al., 2021), scipy 1.7.3 (Virtanen et al., 2020), statsmodels 0.13.2 (Seabold & Perktold, 2010), and pingouin 0.5.3 (Vallat, 2018). ML was performed in python 3.10 using scikit-learn 1.0.2 (Pedregosa et al., 2011) and XGBoost 1.7.1 (T. Chen & Guestrin, 2016).

Sample and Demographics

A total of 62 participants were observed resulting in 50 complete samples with 24 individuals in Group A (3D printers) and 26 individuals in Group B (washing powders). Regarding occupation, 49

of these 50 participants were students, and one was a university staff member. Among the 12 discarded observations, one had to be excluded because of a recording interruption of the eye tracker during the decision phase. Another observation was excluded because the eye tracker was not able to calibrate, most likely due to a facial asymmetry of the participant. The remaining ten discarded observations had to be excluded due to ECG recording issues, particularly because of Bluetooth connection issues between the ECG transmitter and the host computer. The mean age of the remaining 50 participants (29 female and 21 male) was 24.5 years ($SD = 4.9$). Their average participation compensation amounted to €13.5 ($SD = 0.8$).

Correlation of Neurophysiological Features

We investigated correlations of ET and ECG metrics across different time windows for the different experimental periods. As expected, there were no significant correlations between the two sensors. The visualizations for the fixation duration over the different time windows are shown in Supplemental Figure 7 on top. Shorter intervals naturally show correlations with longer ones that comprise them. For example, the time windows from Second 0 to 3 and from 3 to 6 overlap to a large extent with the window from 0 to 5. This results in the red lines of high correlation in the fixation duration plot. While ET features were calculated for the interval lengths (3, 5, 7, 10, 15, 30), the ECG features only comprised the 30-s interval because shorter time windows would have been impractical for most HRV-based features. The bottom part of Supplemental Figure 7 shows the correlations between the HRV features for this interval.

Attitude Towards Salespersons

In order to rule out possible confounds that could arise from different general attitudes toward salespersons, we asked the participants several questions before the actual purchase decision. The internal consistency of the general salesperson attitude scale was acceptable (Tavakol & Dennick, 2011), measured by Cronbach's $\alpha = .76$, and the mean rating was 4 ($SD = 1$), where a high rating corresponds with a high desire to interact with salespersons in general. A Shapiro–Wilk test indicated that the distribution of the mean rating did not depart significantly from normality ($W = 0.98$, $p = .73$), a Bartlett test indicated homoscedasticity ($T = 0.17$, $p = .68$), and a two-sample independent t test did not indicate different means between the groups ($t = 0.7$, $p = .49$). Correlations with personality traits were determined via the Big Five Index–10 scale (Rammstedt et al., 2013). It is plausible that agreeableness is significantly positively correlated ($r = 0.33$, $p = .02$) with a high desire to interact with salespersons.

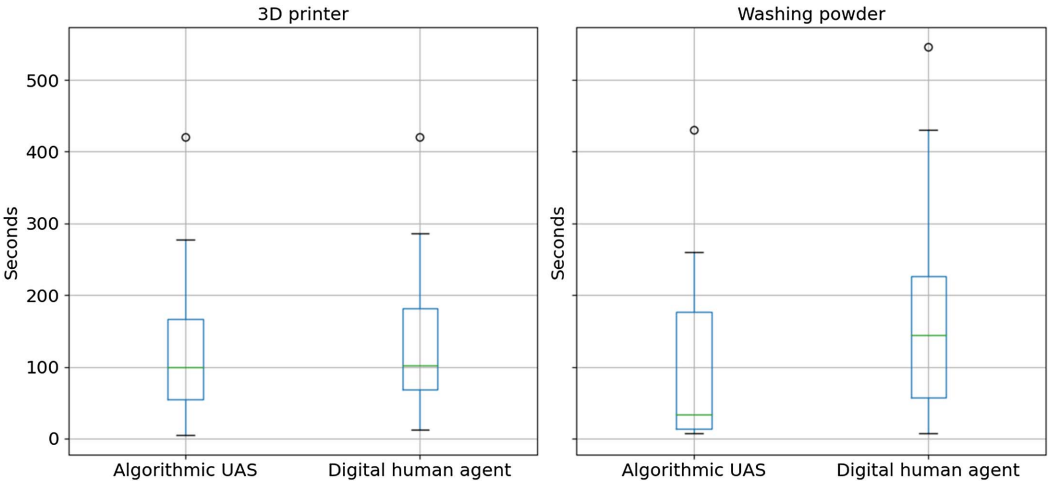
Purchase Duration

The mean purchase duration (from pressing the start button to confirming the purchase) was 247.2 s ($SD = 117.1$) in total, and a normal distribution could not be assumed ($W = 0.94$, $p = .02$). The mean purchase time categorized by groups was 191.3 s ($SD = 85.1$) in Group A (3D printer) and 298.7 s ($SD = 120.2$) in Group B (washing powder). A Mann–Whitney U test indicated a significant difference between the groups ($U = 142.5$, $p < .01$). We see a reason for this difference in the fact that many participants interacted directly with the washing powder packages and regarded the product packages from all sides. For the printer decision, participants pressed the print button but rarely interacted with printed objects because they could visually judge the print quality without touching the objects.

RQ1: Desired Help-Timing

We asked participants about (a) the desired help-timing for an algorithmic UAS and (b) the desired help-timing for a digital human agent. As shown in Figure 5, the early appearance of the algorithmic UAS was particularly relevant to the fast-moving consumer good (FMCG). Reported mean values amounted to 125.5 s ($SD = 113.2$) for both help types (a and b) combined, 103.1 s ($SD = 107.1$) for (a), and 148 s ($SD = 115.8$) for (b). All values related to the duration after activating the start button, which the participants pressed after memorizing the decision requirements on the blackboard. The mean difference (a) − (b) for desired help-timing between the two help providers (desired UAS timing − desired agent timing) was −44.9 s ($SD = 123.3$) for both groups, −12.3 s ($SD = 81.4$) for Group A and −75 s ($SD = 147.4$) for Group B. Multiple Wilcoxon signed rank tests (Wilcoxon, 1992) for paired

Figure 5
Desired Help-Timing for Algorithmic UAS and Digital Human Consultant Stratified by Groups



Note. UAS = user assistance systems. See the online article for the color version of this figure.

samples indicated that the difference (a) − (b) for both product categories ($W = 245$, $p = .02$) and the difference (a) − (b) for Group B ($W = 39.5$, $p = .02$) were significant, while the difference (a) − (b) for Group A was not significant. Participants wanted help from an algorithmic UAS earlier than from a digital human agent, but this was mainly driven by the responses in Group B (washing powder). Overall, the differences in desired help-timing showed the importance of investigating different product categories.

Regarding the most popular choices for algorithmic user assistance, 10 participants in Group A wished for reviews from other consumers, and 14 participants in Group B wished for a product comparison matrix. Hiding irrelevant products and product feature highlighting were the least appreciated help types in both groups. Supplemental Figure 8 shows the complete distribution of the desired help types for algorithmic user assistance.

RQ2: Influence of Knowledge on Help-Timing

Internal consistency of the measured product knowledge items amounted to $\alpha = .76$, which can be seen as acceptable (Tavakol & Dennick, 2011). For the aggregated product knowledge measure, normal distribution and homoscedasticity could be assumed. It amounted to 2.9 ($SD = 1.4$) for Group A, 4.2 ($SD = 1.1$) for Group B, and it significantly differed between the groups ($t = -3.37$, $p < .01$).

As a further control variable, we measured the participants' product involvement. For the respective items, Cronbach's $\alpha = .9$ indicated a very good consistency. Normal distribution and homoscedasticity could be assumed. The mean product involvement of 2.9 ($SD = 1.4$) for Group A, and 4.2 ($SD = 1.1$) for Group B was not significantly different between the product categories ($t = 0.25$, $p = .8$). We expected such a similar product involvement for the different products, due to the equality in monetary incentivization for both groups.

Three ordinary least squares (OLS) regression analyses provided further insight into whether product knowledge influenced desired help-timings for different help providers. First, we considered only product knowledge and product category as independent variables and the absolute desired help-timings as dependent variables (two separate OLS models for algorithmic UAS and digital human agent). For both help types, product knowledge had no significant linear association with desired help-timings. Next, we investigated the same independent variables but used the difference between the desired help-timings as a dependent variable (algorithmic UAS help-timing − digital human agent help-timing). The respective OLS model showed that there was also no significant linear association between product knowledge and the

difference in desired help-timings. Finally, as a robustness check, we included our control variables and compared all three OLS models (desired help-timing for the algorithmic UAS, digital human agent, and the timing difference between the two providers; see Table 2).

In all three constellations, there was no significant linear relationship between product knowledge and desired help-timing. However, we did find a significant linear relationship between participants' openness and their desired help-timing for an algorithmic UAS. Moreover, participants' age and extraversion showed significant linear associations with the desired help-timing for a digital human agent. For the model that accounted for the timing difference between the help providers, the variables age, extraversion, and product involvement showed a significant linear association with the dependent variable.

Overall, we found no support for the influence of product knowledge on desired help-timing. Instead, the OLS models suggested that age, personality traits, and product involvement influence desired help-timing.

RQ3: Cognitive Load Classification

Task Difficulties

We quantified the task difficulty of the generic CL tasks by counting the correct trials for each difficulty level (easy, medium, and hard). The correct completion rates were 146 out of 150 (97.3%) for the easy task, 131 out of 150 (87.3%) for the medium task, and 45 out of 150 (30%) for the hard task and a Kruskal–Wallis test indicated a significant difference between the medians ($H = 99.03$, $p < .01$). Using the NASA TLX questionnaire (Hart, 2006), we measured how demanding our participants perceived the CL tasks and the purchase decision. Regarding the overall task load, a normal distribution could not be assumed for the easy task and the purchase decision (see Supplemental Table 9). Therefore, we conducted a Kruskal–Wallis test that indicated significant differences between the three multitasking difficulty medians ($H = 65.14$, $p < .001$). Yet, due to the rather low internal consistency of the NASA TLX items (Cronbach's $\alpha < .7$, see Supplemental Table 9), we considered only the single item concerning mental strain for further analyses (see Supplemental Table 10). This single item also differed significantly between the tasks ($H = 83.73$, $p < .001$), suggesting that the three CL tasks evoked the desired low, medium, and high CL levels. Next, we tested which of the three CL task difficulty levels was most comparable to the purchase decision task. Pairwise Mann–Whitney U tests indicated significant differences for the tasks with easy and hard difficulty compared to the purchase decision, but this was not the case for the task with medium difficulty (see Supplemental Table 11). The mean perceived task difficulty of the purchase decision was only 0.3 *SD*s less than the

Table 2
OLS Models of Association Between Product Knowledge and Help-Timing

	Model 1: Algorithmic UAS timing			Model 2: Digital human agent timing			Model 3: Timing difference		
Construct	Coefficient	*SE*	*P*	Coefficient	*SE*	*p*	Coefficient	*SE*	*p*
Knowledge	9.14	11.98	.45	10.79	11.6	.358	−1.65	13.26	.902
Involvement	5.47	9.61	.573	−16.64	9.32	.082	22.10	10.65	.045*
Sales representative attitude	−6.73	16.72	.689	−7.71	16.2	.637	0.97	18.52	.958
Agreeableness	11.03	7.76	.164	12.65	7.5	.101	−1.63	8.60	.851
Conscientiousness	9.79	8.76	.271	−0.41	8.5	.962	10.20	9.71	.3
Extraversion	−5.11	6.41	.43	14.82	6.2	.022*	−19.93	7.10	.008**
Openness	−22.12	8.78	.016*	−3.27	8.5	.703	−18.93	9.73	.059
Neuroticism	8.61	7.74	.273	−3.61	7.5	.633	12.23	8.57	.162
Age	3.88	3.34	.252	11.55	3.2	.001**	−7.66	3.70	.045*
Gender (male)	−21.76	37.89	.569	−56.68	36.7	.131	34.92	41.97	.411
Group (B)	−39.00	35.33	.277	32.98	34.3	.342	−71.98	39.13	.074
Intercept	−13.51	167.85	.936	−199.01	162.7	.229	185.50	185.91	.325
R^2		.30			.44			.354	

Note. OLS = ordinary least squares; UAS = user assistance systems; *SE* = standard error.
* $p < .05$. ** $p < .01$.

perceived medium task difficulty. Looking additionally at the boxplots in Supplemental Figure 9, we interpret that, among the available options, the perceived difficulty of the purchase decision can best be matched to the perceived difficulty of the medium task. As a robustness check, we investigated the differences in perceived mental difficulty regarding the purchase decision between the groups. While the perceived mental difficulty in Group B exhibited less variance compared to Group A, we must assume equal mean difficulty between the groups, tested with a Mann–Whitney U test ($U = 368$, $p = .27$).

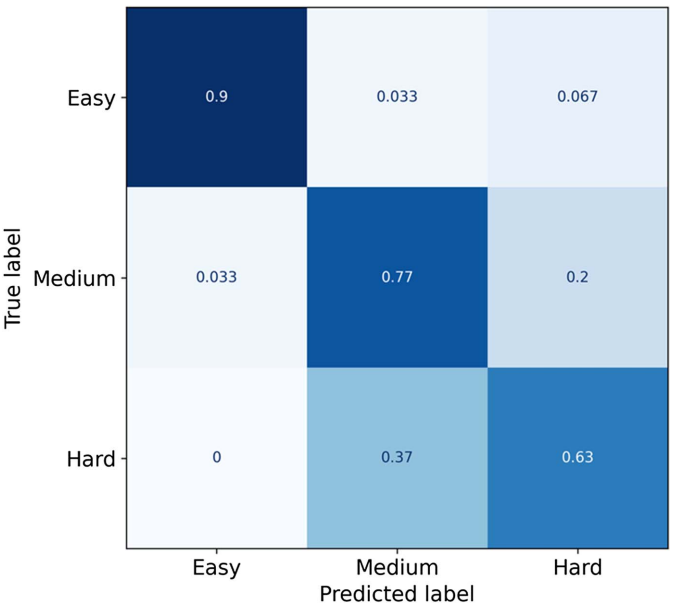
Machine Learning Model

To classify the CL tasks and desired help-timings, we chose an 80% training and 20% testing split method. Instead of selecting a dedicated validation set, we applied a fourfold stratified cross-validation on the training set (Browne, 2000). The optimization metric for classification was accuracy, while (multiclass) negative log-likelihood served as the loss function. Supplemental Table 5 shows the complete hyperparameter space. We used a randomized search approach on the hyperparameters to perform a lightweight tuning, limited to a maximum of 100 iterations. To interpret the feature importance, we used SHapley Additive exPlanations values (Lundberg & Lee, 2017).

First, we solely investigated the generic multitasking difficulty levels. All participants performed three easy trials, three medium trials, and three hard trials for a duration of 30 s each. The best XGBoost model yielded a classification accuracy of .77 for the test set. This means that based on the ET and ECG measurements, we were able to predict with 77% accuracy whether a participant was performing the easy, medium, or hard task. Figure 6 shows the corresponding confusion matrix.

The easy task was classified with a high accuracy of .9, while the medium and hard tasks were not as clearly separable. Despite a correct classification rate of .77 for the medium and .63 for the hard CL levels, these tasks were frequently mutually misclassified. Nonetheless, the classification rates for these two classes were still clearly better than random guesses. A possible explanation for the misclassification between the medium and the hard tasks is the fact that 70% of the participants were unable to successfully complete the hard tasks. We observed that some participants only tracked two elements (the moving balls and appearing numbers) and ignored the additional spinning logo. Even

Figure 6
Confusion Matrix for Best Multitasking Classification Model



Note. See the online article for the color version of this figure.

though this strategy almost certainly resulted in an incorrect answer and no performance-based compensation for the respective round.

The mean absolute SHapley Additive exPlanations values, as shown in Supplemental Figure 10, represent the 20 most important features regarding the multitasking trials in the test set. Different saccade duration and angular speed-related features were prominent (15 of the 20 most important features). This means the required time to jump between AOIs was most discriminative for the CL tasks. Overall, the most important feature was the saccadic mean duration for the whole 30-s period. The number of uniquely fixated objects also played a role, as three features in this regard were among the 20 most important ones. Two blink-related features and one fixation-related feature were also present among them. Regarding the time window sizes, five features related to (3, 7, 15) second time spans, three features related to 30-s time spans, and two features related to 5-s time spans. In our case, ECG and pupillometry features can be considered less important in discriminating between CL difficulty levels as they were not present among the 20 most important features. The best pupillometry feature was variance-related and ranked in 30th place. For ECG, the best feature was the HRV correlation dimension (HRV CD) for the whole trial duration (Bolea et al., 2014), a nonlinear measure for correlations within the signal that ranked 51st place.

We applied the trained multitasking model to the purchase decisions and considered the intervals $[t - 30; t]$ prior to the indicated help timestamps t. Our intention was to identify the prevailing CL level shortly before help was requested. Choosing the same interval duration of 30 s allowed us to create the features analogously to the generic CL tasks. We classified each of the time spans as having either a low, medium, or high CL level. To compare one help interval with one respective nonhelp interval, we used the interval $[t - 60; t - 30]$ as a nonhelp benchmark. For example, if a participant desired help 2 min after pressing the start button, we considered the data for the interval from timestamp 01:00 to 01:30 as the nonhelp benchmark and the data for the interval from timestamp 01:30 to 02:00 as the desired help-timing period. For the desired timing periods of the algorithmic UAS, the model classified high (78%) and medium (12%) CL levels (see Table 3 for absolute counts and Table 4 for classification probabilities). In comparison, most of the nonhelp benchmark intervals (96%) were classified as low CL levels, and only 4% were classified as high CL levels. For the desired timing of the digital human agent, the model classified 88% of the observations as high and 12% as medium. However, the benchmarks for these observations were also mostly classified as high (76%) and medium (18%), while only one observation (2%) was classified as low CL. This implies a difference in CL (an increase from low to high) during the 60 s before the algorithmic UAS was desired but no change in CL during the 60 s before a digital human agent should appear.

Table 3
Classification Results for Machine Learning Model (Match Help Time Spans to Cognitive Load Levels)

Help type	Low	Medium	High
Algorithmic UAS	0	11	39
Algorithmic UAS benchmark	48	0	2
Digital human agent	0	6	44
Digital human agent benchmark	1	11	38

Note. UAS = user assistance systems.

Discussion

For our first research question, relevant insights emerged from the statistical analysis. We found that participants want help earlier from an algorithmic UAS than from a digital human agent. An early appearance of the algorithmic UAS was particularly relevant for the FMCG presented to Group B. The fact that a comparison matrix was the most desired algorithmic help type for the washing powders (see Supplemental Figure 8) suggests that participants were primarily looking for ways to compare the product attributes efficiently. It is likely that they wanted to reduce extraneous CL induced by the rather unfamiliar VR environment. In contrast, when considering the 3D printer decisions, reviews from other consumers were the most desired algorithmic help type. Combined with the insignificant difference in desired help-timing between the algorithmic UAS and the digital human agent when stratifying for Group A, it suggests that these participants were likely seeking help to cope with intrinsic CL.

Reviews were the second most desired help type. As a review by another consumer and an

Table 4
Average Classification Probabilities for Machine Learning Model (Match Help Timespans to Cognitive Load Levels)

Help type	Low *P* (*SD*)	Medium *P* (*SD*)	High *P* (*SD*)
Algorithmic UAS	.05 (0.08)	.31 (0.17)	.65 (0.2)
Algorithmic UAS benchmark	.94 (0.17)	.01 (0.02)	.05 (0.15)
Digital human agent	.03 (0.04)	.28 (0.16)	0.69 (0.18)
Digital human agent benchmark	.05 (0.1)	.31 (0.18)	0.64 (0.2)

Note. UAS = user assistance systems.

expressed opinion by a digital human agent are comparable, we claim that offering a digital human agent as a help provider is more important for the technology product compared to the FMCG. This is further supported by the fact that 3D printers were the product for which our participants reported the least amount of product knowledge. For both groups, participants exhibited a certain reluctance to call for the digital human agent early in the process. A good idea could be to provide a digital human agent as optional help in addition to algorithmic help types which are offered in the first place. Also, when considering nonbinary choices for a certain help offering, our findings clearly highlight the need to customize timing and type of assistance offerings contingent on different scenarios and product categories.

Regarding the second research question, the experiment confirms significant differences in average product knowledge between the technical product and the FMCG. However, we did not find significant linear relationships between product knowledge and desired help-timing for either of the two help providers (and not for the difference in desired help-timing). When controlling for demographics, personality traits, and product involvement, the respective OLS models indicate that participants' age, extraversion, openness, and product involvement have significant linear associations with desired help-timings. The participants' age shows a strong positive linear association with the desired help-timing for a digital human agent ($p = .001$, as shown in Table 2). The positive coefficient indicates that older participants wish to receive help from a digital human agent comparatively late (11.6 s per year). With increasing age, the difference (desired algorithmic UAS timing − desired digital human agent timing) between the desired help-timing also decreases, but this effect is not as strong. Note that the product involvement is not significantly different between the product categories (likely due to the equal monetary incentivization) but displays a positive linear association with the difference between desired timings for the two help providers. More specifically, a one-unit increase on the 7-point Likert scale for product involvement corresponds to a 22.1-s increase in difference. Considering the product involvement coefficient for the timing of the digital human agent ($\beta = -16.64$, $p = .08$), we speculate that as product involvement increases, a digital human agent should appear earlier. To summarize, our OLS models suggest that product knowledge has a subordinate role with respect to desired help-timings. Instead, demographic aspects and personality traits are likely to be more relevant. Product involvement could also play an important role, particularly in scenarios where the variance of product involvement is larger than in ours. In our experiment, we kept the variance in product involvement low by offering the same type of monetary incentive to solve both the 3D printer and the washing powder task.

The analysis of the ML classifications allows us to answer the third research question. Our results suggest that the 30-s periods before the desired help-timings can be mapped with good accuracy to previously determined CL levels, even though the generic tasks were quite different compared to the purchase decisions. This is a promising result, as it suggests that further ML paradigms can potentially be trained with generic CL tasks that are quite different from the actual product decision. Regarding the input features for the XGBoost model, saccade-based metrics were most relevant. Both saccadic angular velocity and saccade duration were highly discriminative. ECG measures were not among the 20 most important

features, which suggests the superiority of the ET sensor over ECG for CL measurement, at least in our relatively brief scenario. As a supplemental data source, ECG can be useful to objectively measure CL, especially over an extended period.

For the 30-s intervals prior to the desired algorithmic UAS help-timing, the ML model predicted medium and high CL levels, but none of the observations were classified as low CL levels. In comparison, the model classified our benchmark interval (60 to 30 s prior to the desired help-timing) mostly as a low CL level. When considering the average class probabilities and their relatively low variances (see Table 4), the benchmark and help intervals exhibit good separability. Overall, an adaptive intervention of an algorithmic UAS, which monitors changes in CL and automatically starts an interaction, seems possible.

Help-timings for a digital human agent were also associated with a medium or high CL level. However, we did not find a significant change in CL levels compared to the respective baselines. The CL level is already medium or high during the baseline interval and does not change when help from a digital human agent is desired. Based on our findings, we argue that the CL level (at this likely later point in the decision-making process) should not be used as the sole indicator to inform a digital human agent about good intervention timing.

Conclusion

This study extends the consumer behavior literature in the emerging subfield of v-commerce. Our statistical analysis investigates the desired help-timings for two different product categories in detail and outlines the need for differentiated treatment. It also reveals behavioral and demographic factors that are linearly associated with desired help-timing. Our study also provides information about the most desired algorithmic help types for different product categories.

Furthermore, we show how ET and ECG data can provide the features for a CL-based ML model, which may benefit the consumer journey. The presented model indicates a good help-timing for an algorithmic UAS while shopping for products or services in a v-commerce context. Even though the ECG measures proved to be supplemental, our study still applies a larger number of ET features than previous studies. For instance, Peukert et al. (2020) used only one ET feature to detect decision phases, and Pfeiffer et al. (2020) limited the number of predictors to four variables at a time.

In the v-commerce context, recognizing and reducing CL is applicable in many ways. Visual and other sensory aids can help to reduce CL and make it easier for consumers to understand information. Moreover, by personalizing a virtual environment, UAS can reduce CL and make it easier for consumers to perform their decision-making processes. CLT can provide 12 principles to break down complex information into smaller, more manageable parts and present it in a clear and concise manner. Our experiment suggests that an ML model can serve as an indicator to invoke an algorithmic UAS which appears just in time and selectively provides the most relevant information to consumers. However, we believe that CL should not be used as the only criterion that determines the current status of the consumer seeking help. Instead, it should be included in multidimensional models to narrow down individualized help time spans for specific environments, products, and situations.

Theoretical Implications

With the proliferation of v-commerce, the emphasis in sales shifts towards providing consumers with a dynamic and interactive shopping experience. This increased attention to consumer experience is driving providers to invest in innovative technology, such as Augmented Reality and VR hardware, and the software to support it. The current rise of AI is likely to accelerate this trend even further, changing the rules for all kinds of retail activities. Our research gives answers to the question by Branca et al. (2023), who asked, "What do we know and what do we not know about consumers' product evaluations in VR?" We complement previous research (a) by showing differences in desired help-timings for different product categories, (b) by identifying relevant impact factors on help-timing, and (c) by applying an extended set of sensors and features in an ML-approach based on CLT. Our results show the feasibility of inferring CL from ET and ECG data, which then serves as a proxy for algorithmic UAS intervention. However, using CL as a single predictor was not sufficient to determine a good point in time for a digital human agent.

Regarding the help type for an algorithmic UAS, our participants requested reviews and

opinions of other consumers most frequently. However, given the fake review problem (He et al., 2022) that currently prevails on several big e-commerce platforms, and combined with the rise of large language models (LLMs), we doubt that written messages or recorded videos will remain as compelling for consumers as they are today. In the second place were side-by-side product comparisons, which outline relevant and detailed information about products in tabular format. Taking CLT and cognitive fit theory (Vessey, 1991) into consideration, such a direct comparison might be feasible for a set of up to four products, which we deem a good maximum comparison capacity. Still, an optimal set size should be the object of further investigations.

The open research questions, such as good intervention timing for digital human agents, require combined efforts, methods, and theories from fields such as economics, neuroscience, and psychology. As sensors like EEG and functional near-infrared spectroscopy (fNIRS) become more precise while steadily shrinking in size and price, collaborative work can help to understand behavioral phenomena in the new context of immersive virtual domains. Applying new combinations of input features and incorporating further psychological effects such as flow (Berger et al., 2023) may also help to explain and model desired help-timing and eventually allow for a better understanding of consumers.

Managerial Implications

We urge practitioners to embrace the challenges and opportunities that new virtual sales channels offer, sometimes even impose. Tech giants are racing for the next breakthrough device after the smartphone, and consumers are wearing an increasing number of sensors that integrate into HMDs and additional wearables, such as wristwatches and earphones. Future shopping assistance will likely involve neurophysiological sensor data, apply ML, and be intelligent. Still, we believe that the human in the loop remains a crucial factor, for instance, as a digital human agent. Although delivered through an avatar, a genuine and actionable recommendation from a real person can still hold more trustworthiness than an automated suggestion from a recommender system (Castelo et al., 2019), particularly in contexts where the user wants that to be seen by others, and to see themselves, as fully human (Heßler et al., 2022). However, LLMs are improving, and a specialized model (in combination with further AI techniques) may soon allow for an intelligent, objective, and thus trustworthy AI sales agent that is perceived as very human-like (Seeger et al., 2021). Revolutionizing real-world call centers and drop-in stores, v-commerce industry pioneers should evaluate how a combination of basic UAS, LLM based AI agents, and digital human agents may provide the most value to the consumer experience.

With respect to ML, our described feature engineering process with different sensors and window sizes may inform how to create an appropriate inference pipeline for help-timing. Our study provides a guideline on how to design a CL-based model that infers desired help-timing for v-commerce consumers. For practitioners with the capability to collect much larger samples than we had, we recommend evaluating time-series-based models. In our approach, we used a small data set, but with more data available, deep time series classifiers like InceptionTime (Ismail Fawaz et al., 2020) or TapNet (X. Zhang et al., 2020) might be suitable models to determine help-timing for a digital human agent. Providers could further combine it with an LLM-based AI agent that has in-depth product knowledge. Overall, such a fine-tuned ML pipeline is likely to enhance consumer experience, increase consumer engagement, and ultimately improve the likelihood of making a sale.

ET has proven to be an accurate sensor that provides both attentional and cognitive metrics. In contrast, we note that the ECG features only had a supplemental character for our study. In a brief period of 30 s, the heart rate is not as informative as the change in pupil dilation or the gaze duration for a certain product. While highlighting the key role of ET, we speculate about the impact of face tracking (FT) in our help-timing prediction endeavors. Realistic synchronization of the cheeks, eyelids, and lips may help to improve the interaction between conversation partners. The next generation of wireless HMDs will integrate ET and FT because good animations and mapping of avatar movements are key in future virtual interactions, not only sales. Thus, incorporating FT seems like a logical next step.

V-commerce providers should consider ethical and privacy-related aspects, as the use of neurophysiological sensors raises many questions. To prevent privacy issues, inference could

be done on the edge device itself, but this would be power-consuming and limited by the embedded processing unit. The European Union enforces special regulatory measures with the AI Act, which could limit online data transfer for inference to a certain degree. However, these regulations are not yet established in detail, and taking influence by means of close cooperation with the regulators seems advisable.

Overall, our article suggests that a virtual showroom is a feasible virtual shopping platform for both FMCG and technical products. Still, we believe that it is not enough to copy prevailing real-world patterns and paradigms into virtual environments. For instance, as space is no constraint in VR, we see a classic shelf arrangement with very low and high product positions as obsolete. Practitioners should put increased effort into identifying and adhering to these new v-commerce rules, such as the need for adjusted ergonomic considerations (Wilson, 1997). Our showroom gives one idea of how a v-commerce sales platform might look, but it is still very close to what is possible in the real world. Engaging VR room designs could go beyond physical limitations and incorporate interesting architectural features. These environments could further incorporate fun games (Tayal et al., 2022) and social activities (Gallace & Girondini, 2022), which might act as an ice-breaker between the consumer and the vendor.

Finally, we advocate for iterative processes when transitioning to virtual sales and help offerings. Our study also describes one part of an iterative research process. Further iterations will introduce the much-spoken-of avatar, and we also plan to evaluate a product comparison matrix UAS for commodity products.

Limitations and Future Research

The limitations of this study can also provide directions and advice for future research. The first concern is the generalizability of the results, as the sample mainly consisted of students. Future research should involve a broader cross-section of society, including different education levels, occupations, and age groups.

Second, future studies should increase the sample size because we were rarely able to assume a normal distribution for statistical testing. For future experiments, it would also make sense to include further product categories (e.g., beverages, food, interior) to obtain a better understanding of product-specific needs. Our results regarding the desired help type also suggest taking a closer investigation of comparison matrices as algorithmic user assistance. A convenient algorithmic UAS for product detail comparison was particularly desired in the FMCG group.

Third, immersion, perceived telepresence, and perceived product involvement could have been increased by adding more sensory channels (particularly audio) to the virtual environment. Future research could mitigate these aspects, for example, by adding sound effects to the products. The room size also had a limiting impact on immersion and telepresence. On several occasions, the experimenter had to interrupt participants and ask them to remain within the defined VR space. Subsequently, they were not able to fully immerse themselves in the virtual space. Future studies with a similar showroom setup should ensure to have at least 25 m^2 of dedicated VR space.

Fourth, the quantitative approach with questionnaires leads to methodical issues like centrality tendencies and questionable consistency, especially for the NASA-TLX items (Hart, 2006). Future studies could mitigate this issue by applying a mixed methods approach and by implementing and validating a more consistent mental difficulty scale.

Fifth, our CL-based ML model predicted help-timing for an algorithmic UAS well but not for a digital human agent. However, we believe that it is feasible to create a predictive model for both help providers. There seem to be other factors for that influence the right intervention timing of digital human agents, which our ET and ECG features do not cover. Furthermore, other model families, such as Hidden Markov Models (Rabiner, 1989) or a deep learning time series classifier, might be able to mitigate the issue and predict timings for both help providers. For a review of different time series classifiers, we refer to Ruiz et al. (2021).

Sixth, future experiments could improve the generic CL tasks or introduce another CL-inducing design, such as a n-back task variant (Jaeggi et al., 2010). We performed the single generic CL task trials sequentially from easy to hard with individually chosen rest periods. Future research could consider a randomized setup with fixed rest periods (which might result in better classification results but bears a risk of reporting

confounds regarding the task order). A broader range of CL tasks could also be considered, for instance, tasks with auditory or haptic components or a classic an n-back task setup. Furthermore, the period of 30 s for the CL tasks is too short for ECG measurements and should be revised for future research. It is also advisable to consider further sensors for CL, such as measuring galvanic skin response and EEG activity, which might be available for future VR devices off-the-shelf.

Future studies may provide deeper insights for the good of both consumers and service providers. New generations of highly immersive VR hardware allow for integrated and appealing experiments. We see the use of neurophysiological sensors in VR as a valuable methodology in experimental consumer behavior research and advocate for further exploration. It remains future work to find indicators for precise help demand prediction regarding a digital human agent. Different age groups and personality traits (like extraversion) may serve as further predictors, as our data has indicated. Incorporating additional neurophysiological aspects, such as emotions (Martinez-Navarro et al., 2019) and stress (Ishaque et al., 2020; Riedl, 2012), is another step to increase the accuracy and generalizability of the ML model. Future research should particularly focus on the prediction of the moment when a digital human (or AI) agent should appear. Most probably, this point in time is more heterogeneously distributed among participants compared to the algorithmic UAS timing.

References

Allen, P. M., Edwards, J. A., Snyder, F. J., Makinson, K. A., & Hamby, D. M. (2014). The effect of cognitive load on decision making with graphically displayed uncertainty information. *Risk Analysis*, *34*(8), 1495–1505. https://doi.org/10.1111/risa.12161

Alzayat, A., & Lee, S. H. M. (2021). Virtual products as an extension of my body: Exploring hedonic and utilitarian shopping value in a virtual reality retail environment. *Journal of Business Research*, *130*, 348–363. https://doi.org/10.1016/j.jbusres.2021.03.017

Anderson, J. R., Bothell, D., & Douglass, S. (2004). Eye movements do not reflect retrieval processes: Limits of the eye-mind hypothesis. *Psychological Science*, *15*(4), 225–231. https://doi.org/10.1111/j.0956-7976.2004.00656.x

Athavipach, C., Pan-Ngum, S., & Israsena, P. (2019). A wearable in-ear EEG device for emotion monitoring. *Sensors*, *19*(18), Article 4014. https://doi.org/10.3390/s19184014

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in Human Behavior*, *22*(4), 685–708. https://doi.org/10.1016/j.chb.2005.12.009

Berger, C., Knierim, M. T., Benke, I., Bartholomeyczik, K., & Weinhardt, C. (2023). *InterFlowCeption: Foundations for technological enhancement of interoception to foster flow states during mental work: About the potential of technologically supported body awareness to promote flow experiences during mental work* [Conference session]. 2023 CHI Conference on Human Factors in Computing Systems, Hamburg, Germany. https://dl.acm.org/doi/abs/10.1145/3544549.3585833

Bigné, E., Llinares, C., & Torrecilla, C. (2016). Elapsed time on first buying triggers brand choices within a category: A virtual reality-based study. *Journal of Business Research*, *69*(4), 1423–1427. https://doi.org/10.1016/j.jbusres.2015.10.119

Bock, O., Baetge, I., & Nicklisch, A. (2014). hroot: Hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120. https://doi.org/10.1016/j.euroecorev.2014.07.003

Bolea, J., Laguna, P., Remartínez, J. M., Rovira, E., Navarro, A., & Bailón, R. (2014). Methodological framework for estimating the correlation dimension in HRV signals. *Computational and Mathematical Methods in Medicine*, *2014*, Article 129248. https://doi.org/10.1155/2014/129248

Bourlakis, M., Papagiannidis, S., & Li, F. (2009). Retail spatial evolution: Paving the way from traditional to metaverse retailing. *Electronic Commerce Research*, *9*(1–2), 135–148. https://doi.org/10.1007/s10660-009-9030-8

Brachten, F., Brünker, F., Frick, N. R. J., Ross, B., & Stieglitz, S. (2020). On the ability of virtual agents to decrease cognitive load: An experimental study. *Information Systems and e-Business Management*, *18*(2), 187–207. https://doi.org/10.1007/s10257-020-00471-7

Branca, G., Marino, V., & Resciniti, R. (2023). How do consumers evaluate products in virtual reality? A literature review for a research agenda. *Spanish Journal of Marketing*. Advance online publication. https://doi.org/10.1108/SJME-07-2022-0153

Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, *44*(1), 108–132. https://doi.org/10.1006/jmps.1999.1279

Caruelle, D., Lervik-Olsen, L., & Gustafsson, A. (2023). The clock is ticking—Or is it? Customer satisfaction response to waiting shorter vs. longer than expected during a service encounter. *Journal of Retailing*, *99*(2), 247–264. https://doi.org/10.1016/j.jretai.2023.03.003

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-dependent algorithm aversion. *JMR, Journal*

of Marketing Research, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Chanel, C. P. C., Wilson, M. D., & Scannella, S. (2019). *Online ECG-based features for cognitive load assessment* [Conference session]. 2019 IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy.

Chen, B., & Yang, D.-N. (2022). *User recommendation in social metaverse with VR* [Conference session]. Proceedings of the 31st ACM International Conference on Information & Knowledge Management, Atlanta, GA, United States.

Chen, T., & Guestrin, C. (2016). *Xgboost: A scalable tree boosting system* [Conference session]. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, United States.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., & Hart, J. C. (1992). The CAVE: Audio visual experience automatic virtual environment. *Communications of the ACM*, *35*(6), 64–72. https://doi.org/10.1145/129888.129892

Cummings, J. J., & Bailenson, J. N. (2016). How immersive is enough? A meta-analysis of the effect of immersive technology on user presence. *Media Psychology*, *19*(2), 272–309. https://doi.org/10.1080/15213269.2015.1015740

Davis, S., Nesbitt, K., & Nalivaiko, E. (2014). *A systematic review of cybersickness* [Conference session]. Proceedings of the 2014 Conference on Interactive Entertainment, Newcastle, NSW, Australia.

Deck, C., & Jahedi, S. (2015). The effect of cognitive load on economic decision making: A survey and new experiments. *European Economic Review*, *78*, 97–119. https://doi.org/10.1016/j.euroecorev.2015.05.004

Dewitte, S., Pandelaere, M., Briers, B., & Warlop, L. (2005). *Cognitive load has negative after effects on consumer decision making*. https://ssrn.com/abstract=813684

Duchowski, A. T. (2017). *Eye tracking methodology: Theory and practice*. Springer. https://doi.org/10.1007/978-3-319-57883-5

Evans, P., & Wurster, T. S. (1999). Getting real about virtual commerce. *Harvard Business Review*, *77*(6), 84–94.

Ferdinando, H., Seppanen, T., & Alasaarela, E. (2016). *Comparing features from ECG pattern and HRV analysis for emotion recognition system* [Conference session]. 2016 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), Chiang Mai, Thailand.

Friemel, C., Morana, S., Pfeiffer, J., & Maedche, A. (2018). On the role of users' cognitive–affective states for user assistance invocation. In F. Davis, R. Riedl, J. vom Brocke, P. M. Léger, & A. Randolph (Eds.), *Information systems and neuroscience* (pp. 37–46). Springer. https://doi.org/10.1007/978-3-319-67431-5_5

Gallace, A., & Girondini, M. (2022). Social touch in virtual reality. *Current Opinion in Behavioral Sciences*, *43*, 249–254. https://doi.org/10.1016/j.cobeha.2021.11.006

Gidlöf, K., Wallin, A., Dewhurst, R., & Holmqvist, K. (2013). Using eye tracking to trace a cognitive process: Gaze behaviour during decision making in a natural environment. *Journal of Eye Movement Research*, *6*, 1–14. https://doi.org/10.16910/jemr.6.1.3

Ginon, B., Stumpf, S., & Jean-Daubias, S. (2016). *Towards the right assistance at the right time for using complex interfaces* [Conference session]. Proceedings of the International Working Conference on Advanced Visual Interfaces, Bari, Italy.

Goy, J.-J. (2013). *Electrocardiography (ECG)*. Bentham Science Publishers. https://doi.org/10.2174/97816080547941130101

Grzyb, T., Dolinski, D., & Kozłowska, A. (2018). Is product placement really worse than traditional commercials? Cognitive load and recalling of advertised brands. *Frontiers in Psychology*, *9*, Article 1519. https://doi.org/10.3389/fpsyg.2018.01519

Guo, G., & Elgendi, M. (2013). A new recommender system for 3D E-commerce: An EEG based approach. *Journal of Advanced Management Science*, *1*(1), 61–65. https://doi.org/10.12720/joams.1.1.61-65

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). *Psycho-physiological measures for assessing cognitive load* [Conference session]. Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Copenhagen, Denmark.

Harris, J. M., Ciorciari, J., & Gountas, J. (2018). Consumer neuroscience for marketing researchers. *Journal of Consumer Behaviour*, *17*(3), 239–252. https://doi.org/10.1002/cb.1710

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. https://doi.org/10.1016/S0166-4115(08)62386-9

Harz, N., Hohenberg, S., & Homburg, C. (2022). Virtual reality in new product development: Insights from prelaunch sales forecasting for durables. *Journal of Marketing*, *86*(3), 157–179. https://doi.org/10.1177/00222429211014902

Häubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science*, *19*(1), 4–21. https://doi.org/10.1287/mksc.19.1.4.15178

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*, *41*(5), 896–921. https://doi.org/10.1287/mksc.2022.1353

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Heßler, P. O., Pfeiffer, J., & Hafenbrädl, S. (2022). When self-humanization leads to algorithm aversion. *Business & Information Systems Engineering*, *64*(3), 275–292. https://doi.org/10.1007/s12599-022-00754-y

Holmqvist, K., Nyström, M., Andersson, R., Dewhurst, R., Jarodzka, H., & van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. OUP Oxford.

Huang, J., Zhao, P., & Wan, X. (2021). From brain variations to individual differences in the color-flavor incongruency effect: A combined virtual reality and resting-state fMRI study. *Journal of Business Research*, *123*, 604–612. https://doi.org/10.1016/j.jbusres.2020.10.031

Hughes, A. M., Hancock, G. M., Marlow, S. L., Stowers, K., & Salas, E. (2019). cardiac measures of cognitive workload: A meta-analysis. *Human Factors*, *61*(3), 393–414. https://doi.org/10.1177/0018720819830553

Ishaque, S., Rueda, A., Nguyen, B., Khan, N., & Krishnan, S. (2020). *Physiological signal analysis and classification of stress from virtual reality video game* [Conference session]. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)Montreal, QC, Canada.

Ismail Fawaz, H., Lucas, B., Forestier, G., Pelletier, C., Schmidt, D. F., Weber, J., Webb, G. I., Idoumghar, L., Muller, P.-A., & Petitjean, F. (2020). Inceptiontime: Finding alexnet for time series classification. *Data Mining and Knowledge Discovery*, *34*(6), 1936–1962. https://doi.org/10.1007/s10618-020-00710-y

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. *Memory*, *18*(4), 394–412. https://doi.org/10.1080/09658211003702171

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review*, *87*(4), 329–354. https://doi.org/10.1037/0033-295X.87.4.329

Kahneman, D. (1973). *Attention and effort* (Vol. 1063). Citeseer.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Lee, H., & Seo, S. (2010). A comparison and analysis of usability methods for web evaluation: The relationship between typical usability test and bio-signals characteristics (EEG, ECG). In D. Durling, R. Bousbaci, L. Chen, P. Gauthier, T. Poldma, S. Roworth-Stokes, & E. Stolterman (Eds.), *Design and complexity—DRS International Conference 2010*. https://dl.designresearchsociety.org/drs-conference-papers/drs2010/researchpapers/73

Lee, Y. J., & Dubinsky, A. J. (2017). Consumers' desire to interact with a salesperson during e-shopping: Development of a scale. *International Journal of Retail & Distribution Management*, *45*(1), 20–39. https://www.emerald.com/insight/content/doi/10.1108/IJRDM-04-2016-0058/full/html

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (pp. 4765–4774). Curran Associates.

Makowski, D., Pham, T., Lau, Z. J., Brammer, J. C., Lespinasse, F., Pham, H., Schölzel, C., & Chen, S. H. A. (2021). NeuroKit2: A Python toolbox for neurophysiological signal processing. *Behavior Research Methods*, *53*(4), 1689–1696. https://doi.org/10.3758/s13428-020-01516-y

Martinez-Navarro, J., Bigné, E., Guixeres, J., Alcañiz, M., & Torrecilla, C. (2019). The influence of virtual reality in e-commerce. *Journal of Business Research*, *100*, 475–482. https://doi.org/10.1016/j.jbusres.2018.10.054

Meißner, M., Pfeiffer, J., Peukert, C., Dietrich, H., & Pfeiffer, T. (2020). How virtual reality affects consumer choice. *Journal of Business Research*, *117*, 219–231. https://doi.org/10.1016/j.jbusres.2020.06.004

Meißner, M., Pfeiffer, J., Pfeiffer, T., & Oppewal, H. (2019). Combining virtual reality and mobile eye tracking to provide a naturalistic experimental environment for shopper research. *Journal of Business Research*, *100*, 445–458. https://doi.org/10.1016/j.jbusres.2017.09.028

Meurisch, C., Mihale-Wilson, C. A., Hawlitschek, A., Giger, F., Müller, F., Hinz, O., & Mühlhäuser, M. (2020). Exploring user expectations of proactive AI systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *4*(4), 1–22. https://doi.org/10.1145/3432193

Miles, W. R. (1929). Ocular dominance demonstrated by unconscious sighting. *Journal of Experimental Psychology*, *12*(2), 113–126. https://doi.org/10.1037/h0075694

Miller, G. A. (1956). The magical number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. https://doi.org/10.1037/h0043158

Moghaddasi, M., Marín-Morales, J., Khatri, J., Guixeres, J., Chicchi Giglioli, I. A., & Alcañiz, M. (2021). Recognition of customers' impulsivity from behavioral patterns in virtual reality. *Applied*

Sciences, *11*(10), Article 4399. https://doi.org/10.3390/app11104399

Morana, S., Pfeiffer, J., & Adam, M. T. P. (2020). User assistance for intelligent systems. *Business & Information Systems Engineering*, *62*(3), 189–192. https://doi.org/10.1007/s12599-020-00640-5

Pan, Y., & Zhang, J. Q. (2011). Born unequal: A study of the helpfulness of user-generated product reviews. *Journal of Retailing*, *87*(4), 598–612. https://doi.org/10.1016/j.jretai.2011.05.002

Park, C.-W., & Moon, B.-J. (2003). The relationship between product involvement and product knowledge: Moderating roles of product type and product knowledge type. *Psychology & Marketing*, *20*(11), 977–997. https://doi.org/10.1002/mar.10105

Park, H., & Kim, S. (2023). Do augmented and virtual reality technologies increase consumers' purchase intentions? The role of cognitive elaboration and shopping goals. *Clothing & Textiles Research Journal*, *41*(2), 91–106. https://doi.org/10.1177/0887302X21994287

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://zbmath.org/1280.68189

Peukert, C., Lechner, J., Pfeiffer, J., & Weinhardt, C. (2020). Intelligent invocation: Towards designing context-aware user assistance systems based on real-time eye tracking data analysis. In F. Davis, R. Riedl, J. vom Brocke, P. M. Léger, A. Randolph, & T. Fischer (Eds.), *Information systems and neuroscience* (pp. 73–82). Springer. https://doi.org/10.1007/978-3-030-28144-1_8

Pfeiffer, J. (2011). *Interactive decision aids in e-commerce*. Springer.

Pfeiffer, J., Pfeiffer, T., Meißner, M., & Weiß, E. (2020). Eye-tracking-based classification of information search behavior using machine learning: Evidence from experiments in physical shops and virtual reality shopping environments. *Information Systems Research*, *31*(3), 675–691. https://doi.org/10.1287/isre.2019.0907

Pham, T., Lau, Z. J., Chen, S. H. A., & Makowski, D. (2021). Heart rate variability in psychology: A review of HRV indices and an analysis tutorial. *Sensors*, *21*(12), Article 3998. https://doi.org/10.3390/s21123998

Plass, J. L., Moreno, R., & Brünken, R. (2010). *Cognitive load theory*. Cambridge University Press. https://doi.org/10.1017/CBO9780511844744

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, *77*(2), 257–286. https://doi.org/10.1109/5.18626

Rammstedt, B., Kemper, C. J., Klein, M. C., Beierlein, C., & Kovaleva, A. (2013). A short scale for assessing the big five dimensions of personality: 10 item big five inventory (BFI-10). *Methods, Data, Analyses*, *7*(2). https://doi.org/10.12758/mda.2013.013

Raut, A., Tiwari, A., Das, S., Saha, S., Maitra, A., Ramnani, R., & Sengupta, S. (2023). Reinforcing personalized persuasion in task-oriented virtual sales assistant. *PLOS ONE*, *18*(1), Article e0275750. https://doi.org/10.1371/journal.pone.0275750

Reinartz, W., Wiegand, N., & Imschloss, M. (2019). The impact of digital transformation on the retailing value chain. *International Journal of Research in Marketing*, *36*(3), 350–366. https://doi.org/10.1016/j.ijresmar.2018.12.002

Riedl, R. (2012). On the biology of technostress: Literature review and research agenda. *The Data Base for Advances in Information Systems*, *44*(1), 18–55. https://doi.org/10.1145/2436239.2436242

Ruiz, A. P., Flynn, M., Large, J., Middlehurst, M., & Bagnall, A. (2021). The great multivariate time series classification bake off: A review and experimental evaluation of recent algorithmic advances. *Data Mining and Knowledge Discovery*, *35*(2), 401–449. https://doi.org/10.1007/s10618-020-00727-3

Russo, J. E., & Leclerc, F. (1994). An eye-fixation analysis of choice processes for consumer nondurables. *Journal of Consumer Research*, *21*(2), 274–290. https://doi.org/10.1086/209397

Salvucci, D. D., & Goldberg, J. H. (2000). *Identifying fixations and saccades in eye-tracking protocols* [Conference session]. Proceedings of the 2000 Symposium on Eye Tracking Research & Applications, Palm Beach Gardens, FL, United States.

Schmutz, P., Roth, S. P., Seckler, M., & Opwis, K. (2010). Designing product listing pages—Effects on sales and users' cognitive workload. *International Journal of Human–Computer Studies*, *68*(7), 423–431. https://doi.org/10.1016/j.ijhcs.2010.02.001

Schnack, A., Wright, M. J., & Holdershaw, J. L. (2021). Does the locomotion technique matter in an immersive virtual store environment? Comparing motion-tracked walking and instant teleportation. *Journal of Retailing and Consumer Services*, *58*, Article 102266. https://doi.org/10.1016/j.jretconser.2020.102266

Seabold, S., & Perktold, J. (2010). *Statsmodels: Econometric and statistical modeling with python* [Conference session]. Proceedings of the Python in Science Conference, Austin, TX, United States.

Seeger, A.-M., Pfeiffer, J., & Heinzl, A. (2021). Texting with humanlike conversational agents: Designing for anthropomorphism. *Journal of the Association for Information Systems*, *22*(4), 931–967 https://doi.org/10.17705/1jais.00685

Seitz, J., & Maedche, A. (2022). Biosignal-based recognition of cognitive load: A systematic review of public datasets and classifiers. In F. D. Davis, R. Riedl, J. vom Brocke, P. M. Léger, A. B. Randolph, & G. R. Müller-Putz (Eds.), *Information systems*

and neuroscience. NeuroIS 2022. Lecture notes in information systems and organisation (pp. 35–52). Springer. https://doi.org/10.1007/978-3-031-13064-9_4

Shim, J. P., Warkentin, M., Courtney, J. F., Power, D. J., Sharda, R., & Carlsson, C. (2002). Past, present, and future of decision support technology. *Decision Support Systems*, *33*(2), 111–126. https://doi.org/10.1016/S0167-9236(01)00139-7

Siegel, E. H., Wei, J., Gomes, A., Oliviera, M., Sundaramoorthy, P., Smathers, K., Vankipuram, M., Ghosh, S., Horii, H., & Bailenson, J. (2021). *HP omnicept cognitive load database (HPO-CLD)-developing a multimodal inference engine for detecting real-time mental workload in VR* [Technical report], HP Labs.

Sutherland, I. E. (1965). The ultimate display. *Proceedings of the IFIP Congress*, *2*, 506–508.

Sweller, J. (2011). Cognitive load theory. *Psychology of Learning and Motivation*, *55*, 37–76. https://doi.org/10.1016/B978-0-12-387691-1.00002-8

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International Journal of Medical Education*, *2*, 53–55. https://doi.org/10.5116/ijme.4dfb.8dfd

Tayal, S., Rajagopal, K., & Mahajan, V. (2022). *Virtual reality based metaverse of gamification* [Conference session]. 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India.

Todd, P. A., & Benbasat, I. (1994). The influence of decision aids on choice strategies under conditions of high cognitive load. *IEEE Transactions on Systems, Man, & Cybernetics*, *24*(4), 537–547. https://doi.org/10.1109/21.286376

Vaessen, B. E., Prins, F. J., & Jeuring, J. (2014). University students' achievement goals and help-seeking strategies in an intelligent tutoring system. *Computers & Education*, *72*, 196–208. https://doi.org/10.1016/j.compedu.2013.11.001

Vallat, R. (2018). Pingouin: Statistics in python. *Journal of Open Source Software*, *3*(31), Article 1026. https://doi.org/10.21105/joss.01026

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, *22*(2), 219–240. https://doi.org/10.1111/j.1540-5915.1991.tb00344.x

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … the SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, *62*, 1–10. https://doi.org/10.1016/j.dss.2014.02.007

Weiß, T. (2023). *Consumer Decisions in Virtual Commerce dataset and source code*. https://osf.io/dkt5v

Wilcoxon, F. (1992). *Individual comparisons by ranking methods*. Springer. https://doi.org/10.1007/978-1-4612-4380-9_16

Wilson, J. R. (1997). Virtual environments and ergonomics: Needs and opportunities. *Ergonomics*, *40*(10), 1057–1077. https://doi.org/10.1080/001401397187603

Xiao, B., & Benbasat, I. (2007). E-commerce product recommendation agents: Use, characteristics, and impact. *Mis Quarterly*, *31*(1), 137–209. https://doi.org/10.2307/25148784

Xiong, R., Kong, F., Yang, X., Liu, G., & Wen, W. (2020). Pattern recognition of cognitive load using EEG and ECG signals. *Sensors*, *20*(18), Article 5122. https://doi.org/10.3390/s20185122

Zaichkowsky, J. L. (1985). Measuring the involvement construct. *Journal of Consumer Research*, *12*(3), 341–352. https://doi.org/10.1086/208520

Zhang, H., Lu, Y., Gupta, S., & Zhao, L. (2014). What motivates customers to participate in social commerce? The impact of technological environments and virtual customer experiences. *Information & Management*, *51*(8), 1017–1030. https://doi.org/10.1016/j.im.2014.07.005

Zhang, X., Gao, Y., Lin, J., & Lu, C.-T. (2020). Tapnet: Multivariate time series classification with attentional prototypical network. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(4), 6845–6852. https://doi.org/10.1609/aaai.v34i04.6165

Zhang, Z., Shang, S., Kulkarni, S. R., & Hui, P. (2013). *Improving augmented reality using recommender systems* [Conference session]. Proceedings of the 7th ACM conference on Recommender systems, Hong Kong, China.

Received April 12, 2023
Revision received December 14, 2023
Accepted January 18, 2024 ■