6-18-2022

MEET YOUR NEW COLLE(AI)GUE – EXPLORING THE IMPACT OF HUMAN-AI INTERACTION DESIGNS ON USER PERFORMANCE

Marvin Braun
University of Goettingen, marvin.braun@uni-goettingen.de

Maike Greve
University of Goettingen, maike.greve@uni-goettingen.de

Johannes Riquel
University of Goettingen, johannes@riquel.de

Alfred Benedikt Brendel
Technisch Universität Dresden, Alfred_benedikt.brendel@tu-dresden.de

Lutz Kolbe
University of Göttingen, lkolbe@uni-goettingen.de

MEET YOUR NEW COLLE(AI)GUE – EXPLORING THE IMPACT OF HUMAN-AI INTERACTION DESIGNS ON USER PERFORMANCE

Research Paper

Braun, Marvin, University of Göttingen, Göttingen, Germany, marvin.braun@uni-goettingen.de

Greve, Maike, University of Göttingen, Göttingen, Germany, maike.greve@uni-goettingen.de

Riquel, Johannes, University of Göttingen, Göttingen, Germany, johannes.riquel@uni-goettingen.de

Brendel, Alfred Benedikt, Technische Universität Dresden, Dresden, Germany, alfred_benedikt.brendel@tu-dresden.de

Abstract

Artificial Intelligence (AI) has an increasing impact on industries, establishing a new way of solving tasks and automating work routines. While AI-based systems have become new colleagues for some processes, the tasks of some humans have shifted towards supervising AI. Essentially, humans need to adapt to a new form of interaction with AI-based systems because AI functioning is more similar to cognitive processes of humans than traditional information systems, e.g., in terms of their intransparent decision making. Previous research indicates that AI adds new challenges to human-computer interaction, and new frameworks for human-AI interaction are developed. However, current research lacks empirical research on the design of such interactions. We conducted a 2x2x2 experiment of AI-supported information extraction and measured the ability of participants to validate the extracted information by the AI. Our results indicate that the design of human-AI interaction significantly impacts users' supervising performance.

Keywords: human-AI interaction, AI-based systems, user performance, signal detection theory

1 Introduction

Through the advances made in AI technology during the last decade, AI has taken over more and more tasks previously executed by humans (Brynjolfsson and Mitchell, 2017; Asan et al., 2020; Lai et al., 2021). It is difficult to define what AI is, but following Brendel et al. (2021), AI is what is currently understood as the most advanced group of self-learning algorithms. Currently, the technology is primarily employed for tasks that involve recognizing patterns, for example, on a visual (object and text recognition) or auditive (speech recognition) level (Amershi et al., 2019). AI uses different technology stacks such as machine learning (ML), natural language processing (NLP), robotics, and computer vision, depending on the use case and the information that shall be processed (Fukas et al., 2021). Meanwhile, AI is considered as a general-purpose technology, underpinning its importance as an innovation driver, and is increasingly embedded into different tasks that are suitable for AI (Brynjolfsson and Mitchell, 2017). The resulting opportunities and potential benefits of AI-based systems are investigated in multiple sectors, for example, in medicine (Hamet and Tremblay, 2017), e.g., where deep learning networks apply computer vision in areas such as cardiology or pathology (Esteva et al., 2021), supply-chain (Nissen and Sengupta, 2006), e.g., in the field of predictive maintenance (Carvalho et al., 2019) and automotive, e.g., through self-driving cars (Badue et al., 2021).

Besides the technological improvements of AI, the interaction between humans and AI is a critical factor for its commercial adaption. According to Deloitte's Global Human Capital Trends, about 60% of institutions plan to use AI as assistance for humans instead of a replacement (Mallon et al., 2020). Therefore, research is needed on how humans and AI can form a team and ultimately solve tasks in cooperation (Lai et al., 2021). The scientific community recognized this shift towards cooperation between humans and AI (Rai et al., 2019), and this implies that human users will need to understand and control AI and its outputs. In comparison to traditional IS, inferences and results presented by AI are not based on clear rules or static code that can be easily understood by humans (Amershi et al., 2019). In the end, results provided by AI are generally based on statistical methods that will most likely never achieve an accuracy of 100% (Brynjolfsson and Mitchell, 2017). Thus, AI regularly produces errors (Amershi et al., 2019) that are often unpredictable (Yang et al., 2020). These unpredictable errors can lead to severe consequences in environments where sensitive or critical information is processed. For example, in the healthcare domain, wrong decisions can directly impact a patient's health (Holzinger et al., 2008).

Previous research has investigated general parameters of human-AI interaction and outlined differences to traditional human-computer interaction (HCI) (Rzepka and Berger, 2018). The consensus is that there is a lack of understanding of how AI-based systems can be effectively designed for workplace implementations (Seiffer et al., 2021). Moreover, large parts of general research related to AI focus on explainable AI (XAI), i.e., making decisions of AI algorithms transparent for humans and, thereby, unraveling the black box. However, the interaction of AI-based systems and users lacks investigation, for instance, how information should be presented to the users and how it results in improved performance. Following Sturm and Peters (2020), users' performance while interacting with AI remains a critical success factor for organizations since AI can often not perform tasks autonomously and a controlling instance is even legally required (Montavon et al., 2018). Thus, in the context of human-AI interaction, we focus on the ability of humans to detect errors of AI-based systems (which we refer to as *AI supervision*) and validate their performance, hereafter referred to as *user performance*. Against this background, this paper aims to answer the following research question:

> ***RQ:*** *How does the information design influence user performance when supervising AI-based systems?*

Based on the general human-AI framework and the signal detection theory (SDT) (Swets and Green, 1963), we derive a research model to investigate the impact of information designs on user performance when interacting with AI. Drawing on the SDT, we define information designs as 'influencing the visual component of information.' Our study considered the following influencing factors for information designs: performance, transparency, and guidance. We derived these by synthesizing the current state of research in the general field of human-AI interaction and combining them with the SDT. We then designed and conducted a 2 (low and high AI performance) x 2 (low and high transparency of results) x 2 (low and high guidance of the user) online experiment. In this experiment, we manipulated the designs for a fixed task to investigate the impact of different treatment configurations on user performance. During the experiment, participants were set into the scenario of supervising an AI-based decision support system (DSS) for information extraction. Users had to either accept the presented results of the AI as correct or mark them as incorrect (AI supervision). The experiment indicates that different design variations influence user performance in the supervision task. Moreover, the results show that the SDT serves as a theoretical foundation for explaining task solutions in human-AI interaction.

2 Theoretical Background

The following section presents relevant background information for measuring and evaluating user performance in human-AI interaction. First, we give a brief overview of the history of human-AI interaction and the current state of research. Second, the SDT is explained and applied to derive hypotheses regarding human-AI interaction.

2.1 Human-AI Interaction

The concept of human-AI interaction (also referred to as human-AI collaboration), which is closely linked to HCI, focuses on the interaction between humans and AI that cooperate to achieve a goal (Sturm et al., 2021). Following the presented HCI framework of Zhang et al. (2002), the human (*user*) wants to perform a *task* that involves (AI-based) information technology (*system*). All three entities (user, task, and system) are shaped by their different characteristics that create and specify the interaction (e.g., user characteristics, the context of the task, and system capabilities). The interaction—the users' contact with the system for conducting a task—then creates outcomes, e.g., achieving the desired goal (Zhang and Li, 2004).

Following recent research, this interaction changes when AI-based systems are employed (Shrestha et al., 2019; Berente et al., 2021). AI creates a new tier of IS that fundamentally reshapes organizational relationships between users and systems (Berente et al., 2021). With the new capabilities of AI, research suggests that teams consisting of humans and AI can be beneficial for task completion by utilizing the individual strengths of both entities. While humans are creative, have empathy, and can flexibly adapt to changing environments and tasks, AI can recognize patterns in data that are not visible for humans and can effectively process information much faster than humans (Dellermann et al., 2019). Depending on the user, AI, and task, human and AI can form different teamwork patterns, where either the human or the AI is in the lead and is supported by the other one (Dellermann et al., 2019; Lubars and Tan, 2019; Shrestha et al., 2019). Choosing the correct mode of human-AI interaction also impacts the task outcome (i.e., the achieved performance) (Dellermann et al., 2019; Sturm and Peters, 2020). Fügener et al. (2021) demonstrated in their multi-method study that human-AI teams can improve the overall task performance, but at the same time, humans are also likely to lose unique human knowledge, which can have other negative impacts.

Since AI is constantly evolving, various other challenges around human-AI teams are actively emerging and researched. Berente et al. (2021) investigate how AI impacts organizations, consequently changing managerial decisions and relationships inside organizations. Teodorescu et al. (2021) show that for AI to achieve a high degree of fair decisions, humans need to augment and correct the machine's decision. Moreover, the authors underline the importance of researching the different aspects of humans augmenting AI (Teodorescu et al., 2021). Furthermore, Kießling et al. (2021) empirically investigate the impact on users if they received information from an AI versus humans (algorithm aversion) and show that higher transparency (more detailed information) does not necessarily increase the user’s trust. Moreover, human-AI teams are researched in different areas, such as service contexts involving conversational agents (Lichtenberg et al., 2021; Riquel et al., 2021), human-in-the-loop approaches in medicine (Kieseberg et al., 2016), or teams of human robots in a manufacturing context (Cimini et al., 2020).

To the best of our knowledge, we are the first to empirically measure the effect of different information designs (interfaces) on user performance in the context of AI supervision. While it is challenging to create an experiment that can generate general findings on human-AI interaction (due to the high diversity of AI applications), we argue that empirical experiments can create insights into the highly discussed field of human-AI teams working on a task together.

User Performance in AI Supervision

The AI, the user, and the task shape the interaction process and thus impact the produced outcome, i.e., affect the user performance during AI supervision. However, it is unclear how and to what extent user performance can be influenced through different information designs. One way to supervise AI is to control if the presented results of the algorithm are correct (human-in-the-loop (Dellermann et al., 2019)). For the task of information extraction, the supervision is accomplished by comparing the *input* that is fed into the AI and the resulting *output*—for example, comparing a scanned textual document (input) and its extracted textual information (output). Visual processing components influence the performance of this supervision task. While user characteristics and the task are usually provided or given by the individual context and hence, unchangeable, the system component allows configurational design changes, which support the visual processing.

The process of visual processing is covered by the SDT of Verghese (2001) as well as by Wolfe and Horowitz (2004), who apply the theory to visual attention (Wolfe and Horowitz, 2004). A visual search task is defined as a setting where an actor looks for a specific item among other items, which distracts the searching person; for example, searching for fitting socks in the laundry (Wolfe and Horowitz, 2004). Visual attention is defined as a signal that uses "the visual system as a stimulus attribute, excluding other input as noise" (Wolfe and Horowitz, 2004, p. 1). Moreover, the authors find multiple visual factors that guide human attention: color, motion, orientation, and size of objects (see Figure 1).

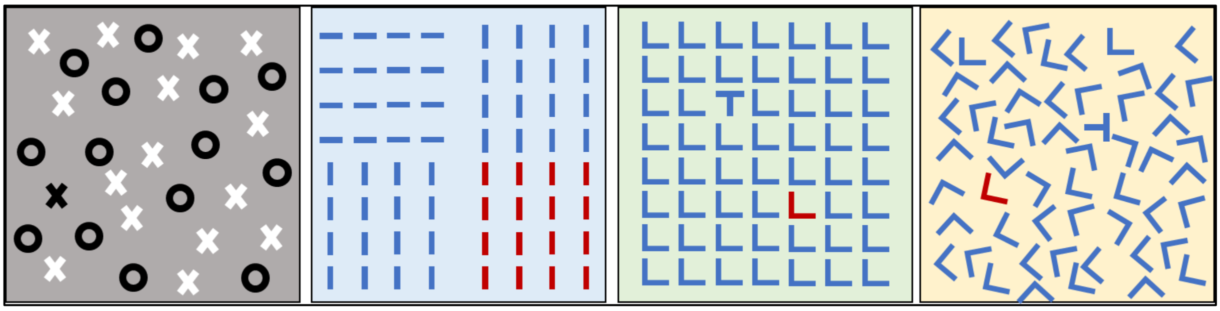


Figure 1. Examples of Factors Guiding Visual Attention in Conjunction: Color, Motion, Orientation, and Size (Wolfe and Horowitz, 2004).

Combining the findings of the SDT with the components of the human-AI interaction reveals hypothetical impacts on user performance for AI supervision tasks. We consider the following visual factors as important for user performance during AI supervision: number of signals, colors of signals, and homogeneity of signals. By matching these three visual factors to the characteristics of human-AI interaction, we define three effects on information design of information that possibly shape user performance (see Table 1).

Information Design	**Definition**	**Related Work**
AI Performance (homogeneity of signals)	The rate at which an AI algorithm produces correct results, i.e., the accuracy of an AI algorithm.	Sturm and Peters, 2020
AI Transparency (colors and shapes of signals)	The rate at which an AI algorithm generates insights about how it came to a specific result, i.e., depicting a probability measure for the result and color indication.	Kieseberg et al., 2016; Wang et al., 2020
AI Guidance (amount of signals)	The rate at which the system guides the user to survey the results of the AI algorithm, i.e., the amount of information which is simultaneously presented to the user.	Kieseberg et al., 2016; Schneeberger et al., 2020

Table 1. Information designs during Human-AI Interaction.

AI performance is indirectly a highly researched topic because it is often considered the primary measure in research for reporting the quality of an AI-based system (Handelman et al., 2019; Lai et al., 2021) and states how well the individual AI algorithm performs its designated task. Moreover, AI Performance influences users' acceptance of the system (Petitgand et al., 2020) and fosters user trust in the system (Wang et al., 2020). Hence, we argue that the performance of AI is an important factor for human-AI collaboration. Despite the high importance of AI performance in general AI research, there is a limited understanding of the empirical impact of AI performance on user performance in the context of human-AI interaction.

AI transparency, which has its origins in the research field XAI, means that the user can comprehend how the AI produced a specific result and gets insights into the working of the algorithm. Transparency is a prerequisite for using AI in different domains such as healthcare due to regulations such as the General Data Protection Regulation introduced by the European Union (Schneeberger et al., 2020). However, transparency is not only required by legal regulations (Amann et al., 2020) but also critical for the human workforce to survey systems (Wang et al., 2020). Moreover, transparency also fosters users' trust in the AI-based system by enabling the comprehension of how an AI derived a specific conclusion (Montavon et al., 2018).

AI guidance is an important aspect of interaction with AI-based systems. When interacting with AI, greater control over the AI seems to reduce the initial reluctance of using such a system (Lockey et al., 2021). Users need to be able to survey the results of the AI (Wang et al., 2018). This is also a legal requirement in the European Union (Schneeberger et al., 2020). To allow a user to control an AI-based system, the user needs the possibility to cooperate and interact with the system instead of just consuming the results of the AI (Holzinger and Kieseberg, 2020).

2.1.1 Impact of AI Performance on User Performance

Building on the SDT and the factors that guide visual attention (see Figure 2), it is estimated that high AI performance leads to decreasing user performance in a task where similar-looking objects need to be compared, such as in the case of AI supervision. Overall, if the AI performs at a high level, fewer extraction mistakes are made. Hence, more similar objects are present, making it harder for users to distinguish between these objects. In conjunction with this, the so-called automation bias needs to be considered a possible effect on user performance (Goddard et al., 2012). The automation bias states that humans tend "to turn over decision processes to automation as much as possible" (Cummings, 2004, p. 2). In this context, humans would not search for any information that falsifies the AI-offered information but rather accept them as they are (Cummings, 2004). Building on the findings of the SDT and the automation bias, we expect that high AI performance increases the cognitive effort that is needed to detect possible mistakes for supervising the AI. Thus, we deduce the following hypothesis:
H1: *The higher the performance of the AI-based software, the lower the user performance in the AI supervision task.*

2.1.2 Impact of AI Transparency on User Performance

The SDT states that the color and size of objects contribute towards the higher visual attention of a human (Wolfe and Horowitz, 2004). In the AI supervision task, higher transparency for the user can be created by highlighting the origin of information (with colors) in the input and linking it to the produced output. This linkage helps users to validate the results of the AI. Thus, it is hypothesized that:
H2: *The higher the transparency of the AI-based software, the higher the user performance in the AI supervision task.*

2.1.3 Impact of AI Guidance on User Performance

According to the SDT, more signals increase the difficulty of recognizing the right signals (Wolfe and Horowitz, 2004). While the task of AI supervision already exercises a certain degree of guidance (guiding the user through the process of supervision), it is hypothesized that when reducing the amount of information (signals) that needs to be supervised by the user at once, the user performance increases. Thus, the following hypothesis is defined:

H3: *The higher the guidance of the AI-based software, the higher the user performance in the AI supervision task.*

3 Methodology

An online experiment with a 2 (AI performance: high vs. low) x 2 (AI transparency: high vs. low) x 2 (AI guidance: high vs. low) factorial design and a quantitative follow-up questionnaire was conducted to examine the effects of the three derived information designs on user performance (see Table 2).

		AI Transparency			
		Low		**High**	
AI Guidance	**Low**	Low AI Performance (n = 22)	High AI Performance (n = 24)	Low AI Performance (n = 22)	High AI Performance (n = 21)
	High	Low AI Performance (n = 24)	High AI Performance (n = 24)	Low AI Performance (n = 25)	High AI Performance (n = 25)

Table 2. Design Treatment Groups of the Experiments.

The experiment scenario is a fictional situation where participants are put into the role of human resources employees who want to extract information from curriculum vitae (CVs) of applicants supported by a new AI-based system. This fictional setting was selected to enable the participants (mostly undergraduate and graduate students in economics, who regularly apply for jobs and internships) to relate to the fictional setting. This was validated via a scenario check (standard deviation (SD) = 6.76, mean (M) = .596).

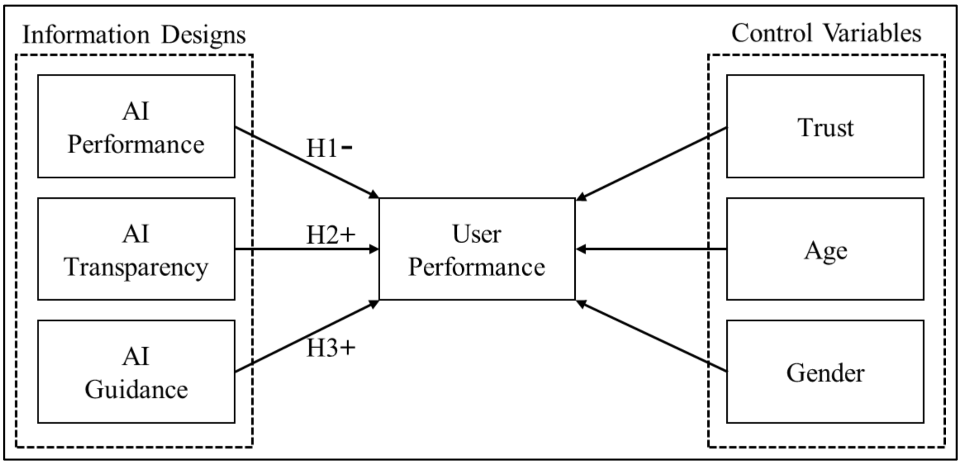


Figure 2. Research Model.

In total, 201 participants took part in the distributed experiment. 13 data sets were removed because participants did not pass the attention checks. One data set was eliminated because its performance was 5%, and the participant's scenario check was 1 on a scale of 1 to 10, where 10 indicates that the participant can put themself into the described situation. Thus, our analysis is based on 187 data sets.

The participants were equally distributed among the eight groups (see Table 2). The participants' ages ranged from 20 to 37 years (M = 25.10 years; SD = 2.51), with 90% students and 10% full and part time employed participants. Moreover, 46.5% of participants were female. The derived hypotheses are embedded into a research model (see Figure 2) that investigates the effect of information designs on user performance.

3.1 Task and Procedure

Before the start of the actual experiment, the participants were introduced into the scenario with the following information provided:

> *"Please imagine that you are a recruiter working in the human resources department. One part of your task is to analyze application documents that are sent to you by applicants. These applications can be for example motivational letters or CVs (Curriculum Vitae).*
>
> *Until now, you had to manually (by hand) transfer the information from these documents into your computer systems. Today, a new software tool was introduced that automates the information extraction process by using artificial intelligence (AI). However, you still need to review the information and check the information for mistakes."*

Followed by the introductory text, the participants were given a simple task to make the procedure easier to understand. Next, the participants were shown five different CVs of fictional applicants. These CVs had different layouts (one of these can be seen in Figure 3) and included fictional information such as their address, contact data, and details about academic and professional careers. The participant's task was to identify errors (misspelled information) made by the AI. After the experiment, the participants were asked to answer a follow-up questionnaire which included latent variables, control variables, attention checks, and demographic questions.

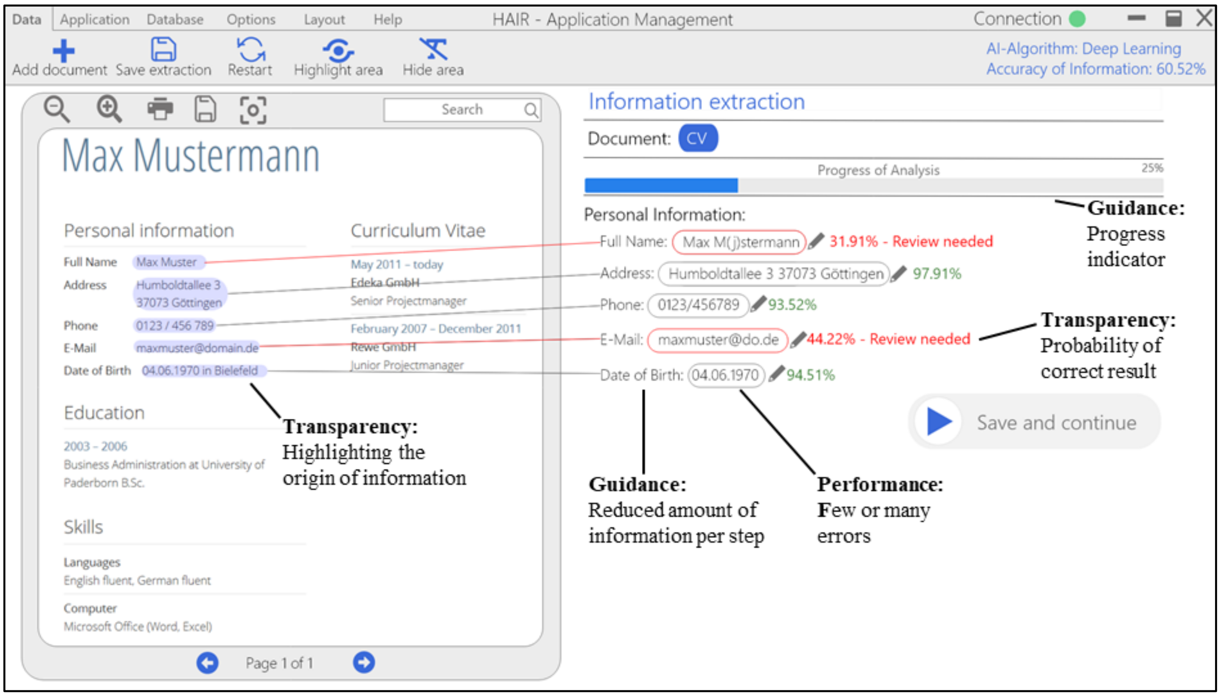
3.2 Design Manipulations

In total, eight different design variations are created and based on the six visual severities of the three information designs. Table 3 depicts the designs and their two forms of visual severity, which provide the basis for the design variations.

Information Design	**Visual Severity**	
	Low	**High**
AI Performance	Many errors.	Few errors.
AI Transparency	No added visual guiding factors.	Origins of information are colored.
AI Guidance	All information is presented at once. No visible progress indicators.	Information is presented stepwise. Progress is indicated by a progress bar.

Table 3. Visual Severities of the Information Designs.

The modularity of the designs and the visual instantiation of our prototype allowed us to change single elements on a very precise level. An example of a design with low AI performance, high AI transparency, and high AI guidance can be seen in Figure 3. The left part of the application remained unchanged for most designs; only high transparency added colors to the origin of the information as well as lines connected to the output. Moreover, for high AI transparency the certainty of the AI about the extracted information is visualized for the user to indicate the probability of the extracted information being correct. Groups that received high AI performance as a treatment had one incorrectly extracted information per CV, resulting in five incorrect pieces of information in total (90% AI performance) that needed to be recognized by the user. In contrast, groups that received low AI performance as a treatment had four incorrect extracted pieces of information per CV, resulting in a total of 20 incorrect pieces of information (60% AI performance). Groups that received the treatment of high AI guidance treatment were guided through the process of AI supervision. In contrast to the groups with low AI guidance, groups with high AI guidance information were presented in batches of 4-6 pieces of information at a time to reduce possible noise in the context of the SDT.


Data
Application
Database
Options
Layout
Help
HAIR - Application Management
Connection
Add document
Save extraction
Restart
Highlight area
Hide area
AI-Algorithm: Deep Learning
Accuracy of Information: 60.52%
Search
Page 1 of 1
Information extraction
Document: CV
Progress of Analysis
25%
Save and continue
Guidance: Progress indicator
Transparency: Probability of correct result
Transparency: Highlighting the origin of information
Guidance: Reduced amount of information per step
Performance: Few or many errors

Max Mustermann

Personal information

Full Name Max Muster
Address Humboldtallee 3
37073 Göttingen
Phone 0123 / 456 789
E-Mail maxmuster@domain.de
Date of Birth 04.06.1970 in Bielefeld

Curriculum Vitae

May 2011 – today
Edeka GmbH
Senior Projectmanager

February 2007 – December 2011
Rewe GmbH
Junior Projectmanager

Education

2003 – 2006
Business Administration at University of Paderborn B.Sc.

Skills

Languages
English fluent, German fluent

Computer
Microsoft Office (Word, Excel)

Personal Information:
Full Name: Max M(j)stermann 31.91% - Review needed
Address: Humboldtallee 3 37073 Göttingen 97.91%
Phone: 0123/456789 93.52%
E-Mail: maxmuster@do.de 44.22% - Review needed
Date of Birth: 04.06.1970 94.51%

Figure 3. AI-based System Design with high AI Transparency, high AI Guidance, and low AI Performance.

3.3 Dependent Variable

As the dependent variable of the experiment, user performance was measured. User performance needs to be considered in the individual task and technology (Rzepka and Berger, 2018; Sturm and Peters, 2020). Nevertheless, we argue that several application areas of AI include similar tasks where errors made by AI need to be recognized through supervision. Following the definition of performance by Eccles and Pyburn (1992), the multidimensionality of the construct of performance is apparent. However, in the context of this study, we specify user performance as the sole recognition of AI errors by humans. Hence, we define user performance in the following way:

$$\boldsymbol{User\ Performance = \frac{\sum Identified\ Wrong\ Information}{\sum Wrong\ Information}}$$

The variable measures the amount of correctly identified wrong information by the participants. We choose this as the main dependent variable because, in a real task of validating the results of an AI algorithm, users would focus on finding incorrectly extracted information. This means that if users find all errors (either 5 or 20), they reach 100% user performance.

3.4 Control Variables

We embedded three control variables to measure possible impacts of these on user performance: age, gender, and trust. Age and gender have been selected to assure that the findings of this work are not limited to a specific group. Moreover, trust in AI has been discussed in the literature as a possible negative influence on user performance (Reddy et al., 2020). Hence, the variable is included as a control variable to be able to exclude its effect on user performance. The items were adapted from Cyr et al.

(2009). The construct was validated by confirmatory factor analysis (CFA) and a calculation of Cronbach's Alpha (α). The CFA was conducted to confirm a proper correlation between single items and the underlying construct trust that was supposed to be measured. The factor loadings of trust range from .892 to .926, indicating a very high internal correlation since they exceed the threshold of .6 (Kline, 2014). Finally, Cronbach's α was calculated for trust to investigate the internal consistency of the construct. Following the proposal of Cortina et al. (1993), we define values greater than .70 as acceptable. We measured a value of .942, indicating a high internal consistency.

4 Results

The data from the proposed research model (see Figure 3) is analyzed by a three-factor Analysis of Variance (ANOVA) in SPSS (Version 26). The effects of the treatment variables on user performance are analyzed first.

Treatment	**df**	**SS**	**F**	***p***
PERF	**1**	**.181**	**12.147**	**.001*** **
TRANSP	**1**	**.109**	**7.298**	**.008** **
GUID	1	.019	1.270	.261
PERF*TRANSP	1	.002	.111	.739
PERF*GUID	**1**	**.075**	**5.009**	**.026***
TRANSP*GUID	1	.016	1.043	.308
PERF*TRANSP*GUID	1	>.001	.000	.986
Residuals	3.076	186		
Significant Codes: *** < .001; ** < .01; * < .05 R Squared = .131				

PERF = AI Performance; TRANSP = AI Transparency; GUID = AI Guidance
df = degrees of freedom, SS = Sum of Squares, F = F-statistic, p = p-value				

Table 4. Three-factorial ANOVA.

The ANOVA examines the effect of AI performance, AI transparency, and AI guidance on user performance (see Table 4). A statistically significant effect on user performance was measured for AI performance (F = 12.147, p = .001) and AI transparency (F = 7.298, p = .008). For AI guidance, no statistically significant main effect was measured (F = 1.270, p = .261). Moreover, a significant interaction effect between AI guidance and AI performance was measured (F = 5.009, p = .026). In the following, we analyze the individual effects. The first significant variable, AI performance, has a negative impact on user performance. The ANOVA shows that participants in groups of low AI performance had an average user performance of M = .938 (SD = .079). In groups where AI performance was high, the following average user performance was measured: M = .877 (SD = .158). These results support hypothesis H1 stating that the AI performance of the AI-based system has a negative impact on user performance.

For AI transparency, a positive effect on user performance was measured. The ANOVA indicates that groups with low AI transparency achieved an average user performance of M = .881 (SD = .129). In comparison, groups with high AI transparency achieved an average user performance of M = .932 (SD = .123). Therefore, the hypothesis that AI transparency enhances user performance (H2) is supported as well. For the treatment of AI guidance, no significant main effect was measured.

The significant interaction effect of AI performance and AI guidance can be seen in Figure 4 (left graph). First, as indicated by the main effect of the AI performance variable, groups with low performance performed better on average. When AI performance is high, user performance decreases; however, if AI guidance is high (M = .906, SD = .147), user performance does not decrease as much as with low AI guidance (M = .844, SD = 164). For the other interactions, AI performance * AI transparency and AI transparency * AI guidance no significant effects were measured (see Figure 4, graph middle and right). Furthermore, no significant interaction effect was found for AI performance*AI transparency*AI guidance.

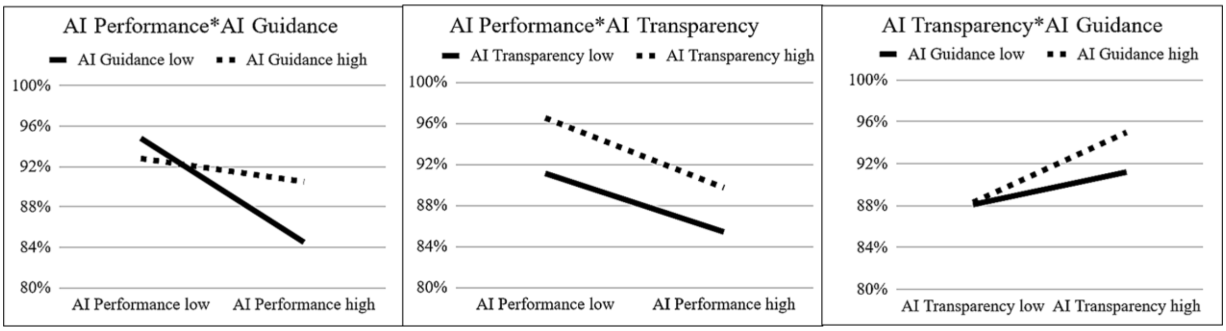


Figure 4. Measured Interaction Effects.

To conclude our results, hypotheses H1 and H2 are supported (see Table 5). H3 is not supported, according to the main effects. However, an interaction effect between AI guidance and AI performance was measured (p = .026).

Hypothesis	Relationship	Support
H1	AI Performance → User Performance	**Yes**
H2	AI Transparency → User Performance	**Yes**
H3	AI Guidance → User Performance	No

Table 5. Results of the Experiment.

5 Discussion

In the following section, we first summarize our findings. Next, we derive theoretical contributions and practical implications from the findings. Finally, the limitations of our study are described, and an agenda for future research is derived.

5.1 Summary of Findings

Our study explored the effect of information designs on user performance during human-AI interaction. The 2x2x2 online experiment included 187 participants who supervised and controlled the results of an AI algorithm and decided if its output was correct or incorrect (supervision task). By utilizing the SDT, we were able to derive explanations for the impact of information designs on user performance. The study results add insights to human-AI interaction and confirm that different presentations of information need to be considered when creating AI-based software.

First, we found that AI performance significantly negatively affects user performance. The explanation for this effect was derived from SDT and the automation bias. We hypothesized that if AI performance is low, the user can recognize differences between input (original document) and output (extracted information) more easily because there are more differences, reducing the cognitive effort needed and thus, lowering the tendency towards an automation bias. In contrast, if AI performance is high, there are only a few differences between input and output, making it more difficult for the user to distinguish between these signals, thus increasing the cognitive effort needed and favoring tendencies towards the automation bias. At first, this finding is contrary to existing research (Fügener et al., 2021); however, we hypothesize that the individual task context (i.e., tasks involving texts versus images) influences the interaction and, thus, the individual user performance. In our task, participants had the chance to directly compare the correct (original) information to the extracted information to AI, whereas in the task presented by Fügener et al. (2021), images are compared against each other leaving room for interpretation. This finding implicates that with increasing AI performance, which is an unavoidable and obvious goal of the development of AI, developers of AI-based systems need to consider mechanisms to support the user when interacting with AI. Our results indicate that high AI performance leads to a decrease in user performance. This suggests that developers should consider visual mechanisms, such as AI guidance and AI transparency, to counter the negative effects of AI performance instead of focusing on AI performance solely. One solution could be reducing the amount of information (high AI guidance) that needs to be processed by the user simultaneously. We find that this significantly reduces the negative effect of AI performance on user performance.

Second, the results highlight the importance of AI transparency in the context of user performance. In our experiment, AI transparency had a significant positive effect on user performance. The SDT states that colors are very strong visual guiding factors that beat other factors such as shapes and motions. In our experiment, colors were used to draw participants' attention and help them compare input and output. AI transparency has been considered an important characteristic in multiple contexts such as legal requirements (Wang et al., 2018; Longo et al., 2020) and explainability (Schneeberger et al., 2020; Lockey et al., 2021) to get insights into the working of AI. The findings indicate that it can also be directly linked to user performance. This result underpins the importance of transparency as a legal requirement and for general human-AI interaction and especially user performance.

Finally, the analysis indicates a significant interaction effect of AI performance and AI guidance. While AI guidance itself was not significant, this effect can be explained with SDT as well. By reducing the amount of information that the user processes at the same time (which is our definition of high AI guidance), high AI performance (less incorrect information that needs to be recognized by the user) is better processable by users because it is easier for them to find differences between the input and output in comparison to low AI guidance. The interaction effect suggests that combinations of different information designs have different impacts on user performance.

5.2 Theoretical Contribution and Practical Implications

This study adds new knowledge to the existing literature of human-AI interaction and provides a theoretical contribution as well as practical implications. To the best of our knowledge, we are the first to empirically investigate the effects of information designs on user performance. The results show that the information designs and their impact on AI-based systems play a crucial role in general human-AI interaction and enable users to conduct (visual-based) tasks in cooperation with AI, such as supervision tasks. The results of our study line up into the growing corpus on investigating human performance during human-AI collaboration (Fügener et al., 2021; Sturm et al., 2021; Teodorescu et al., 2021) and contributes to it by providing empirical evidence on how user performance is influenced through visual components during the supervision of AI.

Moreover, we can extend the understanding of human-AI interaction by showing that the SDT can be utilized to explain detailed effects on user performance when dealing with AI-based systems that involve visual tasks. The SDT is specifically more relevant for human-AI interaction than traditional HCI because of the high uncertainty and the limited explainability of derived results (black box). The SDT provides guidelines on how users' attention is guided to the correct elements and can be used to optimize the design of human-AI interfaces on a visual level.

We also derive several practical implications from our study. Our experiment focused on user performance during interaction with a DSS. Often, these systems are used in healthcare (so-called clinical DSS), where information, in general, has special requirements and can have direct consequences on the patients' health. Our results imply several design guidelines for developers of AI-based DSS, especially for information-sensitive environments, to ensure high information quality. Developers with visual human-AI interfaces should consider mechanisms to guide the user's attention towards important information. Further, our results indicate that developers of AI-based systems should not only employ mechanisms to reduce negative effects of AI performance, but also avoid unnecessary visual designs that distract users' attention.

5.3 Limitations and Future Research

This study has several limitations and provides opportunities for future research. The first limitation is the artificial setting of the experiment, which most participants probably have never encountered before. However, we tried to overcome this limitation by implementing a scenario check. Another limitation is that most of our participants were students. Nevertheless, findings derived from samples of students are considered to hold and can usually be generalized (Compeau et al., 2012). As suggested by several authors, human-AI interaction is also dependent on the individual setting and task (Rzepka and Berger, 2018; Seiffer et al., 2021). Our study peeks at how information designs affect human-AI interaction during an AI supervision setting. The experiment was set in the domain of information extraction, but the application areas of AI are very diverse; thus, the generalizability of our findings is not confirmed. Future research could conceptualize the effect of informative presentations on different tasks and types of human-AI interaction.

Additionally, the concrete implementations of the information designs such as AI transparency are subjective; other designs could influence the results (e.g., using different colors). Future research could investigate how designs can specifically be developed depending on the type of AI and considered tasks.

Another limiting factor is the design of errors in our experiment. It was made relatively easy to find all errors for our participants (indicated by the overall high user performance of M = .907 (SD = .128)). The AI algorithm recognized all information in the experiment, but some were misspelled or completely wrong. Moreover, we acknowledge that the difficulty of detecting different errors of the AI varies. We propose that future research should elaborate on how user performance changes when AI does not recognize all information, forcing users to focus on multiple error types, thus increasing the cognitive effort needed to solve the task successfully. Apart from that, the high AI transparency only displayed low probabilities in case of wrong extraction, which further simplified the detection of incorrect extraction. Future research could examine how users deal with misleading transparency (e.g., green colors for incorrect values, red colors for correct values).

References

Amann, J., A. Blasimme, E. Vayena, D. Frey and V. I. Madai. (2020). "Explainability for artificial intelligence in healthcare: a multidisciplinary perspective." *BMC Medical Informatics and Decision Making 20* (1), 310.

Amershi, S., D. Weld, M. Vorvoreanu, A. Fourney, B. Nushi, P. Collisson, E. Horvitz. (2019). Guidelines for human-AI interaction. In: *Conference on Human Factors in Computing Systems - Proceedings*, p. 13. Association for Computing Machinery.

Asan, O., A. E. Bayrak and A. Choudhury. (2020). "Artificial Intelligence and Human Trust in Healthcare: Focus on Clinicians." *Journal of Medical Internet Research 22* (6), e15154.

Badue, C., R. Guidolini, R. Carneiro, P. Azevedo, V. B. Cardoso, A. Forechi, A. F. De Souza. (2021). "Self-driving cars: A survey." *Expert Systems with Applications 165*, 113816.

Berente, N., J. Recker and R. Santhanam. (2021). "Managing Artificial Intelligence." *MIS Quarterly: Management Information Systems 45* (3), 1433–1450.

Brendel, A. B., M. Mirbabaie, T.-B. Lembcke and L. Hofeditz. (2021). "Ethical Management of Artificial Intelligence." *Sustainability 13* (4), 1974.

Brynjolfsson, E. and T. Mitchell. (2017). "What can machine learning do? Workforce implications: Profound change is coming, but roles for humans remain." *Science 358* (6370), 1530–1534.

Carvalho, T. P., F. Soares, R. Vita, R. da P. Francisco, J. P. Basto and S. G. S. Alcalá. (2019). "A systematic literature review of machine learning methods applied to predictive maintenance." *Computers and Industrial Engineering 137*, 106024.

Cimini, C., F. Pirola, R. Pinto and S. Cavalieri. (2020). "A human-in-the-loop manufacturing control architecture for the next generation of production systems." *Journal of Manufacturing Systems 54*, 258–271.

Compeau, D., B. Marcolin, H. Kelley and C. Higgins. (2012). "Generalizability of information systems research using student subjects A reflection on our practices and recommendations for future research." *Information Systems Research 23* (4), 1093–1109.

Cortina, J. M. (1993). "What is coefficient alpha? An examination of theory and applications." *Journal of Applied Psychology 78* (1), 98–104.

Cummings, M. L. (2004). Automation bias in intelligent time critical decision support systems. In: *Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference,* Vol. 2, pp. 557–562. Reston, Virigina: American Institute of Aeronautics and Astronautics.

Cyr, D., M. Head, H. Larios and B. Pan. (2009). "Exploring human images in website design: A multi-method approach." *MIS Quarterly: Management Information Systems 33* (3), 539–566.

Dellermann, D., P. Ebel, M. Söllner and J. M. Leimeister. (2019). "Hybrid Intelligence." *Business & Information Systems Engineering 61* (5), 637–643.

Eccles, R. G. and P. J. Pyburn. (1992). "Creating a Comprehensive System to Measure Performance - Financial Results Should Not Generate the Most Rewards." *Management Accounting 74* (4), 41–44.

Esteva, A., K. Chou, S. Yeung, N. Naik, A. Madani, A. Mottaghi, R. Socher. (2021). "Deep learning-enabled medical computer vision." *Npj Digital Medicine 4* (1), 5.

Fügener, A., J. Grahl, A. Gupta and W. Ketter. (2021). "Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI." *MIS Quarterly 45* (3), 1527–1556.

Fukas, P., J. Rebstadt, F. Remark and O. Thomas. (2021). Developing an Artifificial Intelligence Maturity Model for Auditing. In: *ECIS 2021 Research Papers*, pp. 6–14. Twenty-Ninth European Conference on Information Systems (ECIS 2021).

Goddard, K., A. Roudsari and J. C. Wyatt. (2012). "Automation bias: A systematic review of frequency, effect mediators, and mitigators." *Journal of the American Medical Informatics Association 19* (1), 121–127.

Hamet, P. and J. Tremblay. (2017). "Artificial intelligence in medicine." *Metabolism: Clinical and Experimental 69*, S36–S40.

Handelman, G. S., H. K. Kok, R. V. Chandra, A. H. Razavi, S. Huang, M. Brooks, H. Asadi. (2019). "Peering into the black box of artificial intelligence: Evaluation metrics of machine learning methods." *American Journal of Roentgenology 212* (1), 38–43.

Holzinger, A., R. Geierhofer, F. Mödritscher and R. Tatzl. (2008). "Semantic information in medical information systems: Utilization of text mining techniques to analyze medical diagnoses." *Journal of Universal Computer Science 14* (22), 3781–3795.

Holzinger, A. and P. Kieseberg. (2020). *Machine Learning and Knowledge Extraction*. (A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl, Eds.)*CD-MAKE: International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, Vol. 12279. Cham: Springer International Publishing.

Kieseberg, P., E. Weippl and A. Holzinger. (2016). "Trust for the doctor-in-the-loop." *ERCIM News 104* (1), 32–33.

Kießling, S., K. Figl and U. Remus. (2021). "Human Experts or Artificial Intelligence? Algorithm Aversion in Fake News Detection." *ECIS 2021 Proceedings Research Papers* 6–14.

Kline, P. (2014). *An Easy Guide to Factor Analysis*. *An Easy Guide to Factor Analysis*. Routledge.

Lai, Y., A. Kankanhalli and D. C. Ong. (2021). Human-AI collaboration in healthcare: A review and research agenda. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2020-Janua, pp. 390–399.

Lichtenberg, S., J. Bührke, A. B. Brendel, S. Trang, S. Diederich and S. Morana. (2021). "Let Us Work Together"– Insights from an Experiment with Conversational Agents on the Relation of Anthropomorphic Design, Dialog Support, and Performance, pp. 299–315.

Lockey, S., N. Gillespie, D. Holm and I. A. Someh. (2021). A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2020-Janua, pp. 5463–5472.

Longo, L., R. Goebel, F. Lecue, P. Kieseberg and A. Holzinger. (2020). Explainable Artificial Intelligence: Concepts, Applications, Research Challenges and Visions, pp. 1–16.

Lubars, B. and C. Tan. (2019). Ask not what AI can do, but what AI should do: Towards a framework of task delegability. In: H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc.

Mallon, D., J. Harris, B. Denny, J. Schwartz and S. Elliott. (2020). Superteams : Putting AI in the group CAPITAL H Superteams : Putting AI in the group. Retrieved from https://www2.deloitte.com/us/en/insights/focus/human-capital-trends/2020/human-ai-collaboration.html

Montavon, G., W. Samek and K. R. Müller. (2018). "Methods for interpreting and understanding deep neural networks." *Digital Signal Processing: A Review Journal 73*, 1–15.

Nissen, M. E. and K. Sengupta. (2006). "Incorporating software agents into supply chains: Experimental investigation with a procurement task." *MIS Quarterly: Management Information Systems 30* (1), 145–166.

Petitgand, C., A. Motulsky, J. L. Denis and C. Régis. (2020). "Investigating the barriers to physician adoption of an artificial intelligence-based decision support system in emergency care: An interpretative qualitative study." *Studies in Health Technology and Informatics 270*, 1001–1005.

Rai, A., P. Constantinides and S. Sarker. (2019). "Next generation digital platforms : toward human-AI hybrids." *MIS Quarterly 44* (1), iii–ix.

Reddy, S., S. Allan, S. Coghlan and P. Cooper. (2020). "A governance model for the application of AI in health care." *Journal of the American Medical Informatics Association : JAMIA 27* (3), 491–497.

Riquel, J., A. B. Brendel, F. Hildebrandt, M. Greve and L. Kolbe. (2021). ""Even the Wisest Machine Makes Errors" – An Experimental Investigation of Human-like Designed and Flawed Conversational Agents." *Proceedings of the 42nd International Conference on Information Systems (ICIS 2021).*

Rzepka, C. and B. Berger. (2018). User Interaction with AI-enabled Systems: A systematic review of IS research. *Proceedings of the 39th International Conference on Information Systems (ICIS 2018).*

Schneeberger, D., K. Stöger and A. Holzinger. (2020). The European Legal Framework for Medical AI. In: A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 12279 LNCS, pp. 209–226. Cham: Springer International Publishing.

Seiffer, A., U. Gnewuch and A. Mädche. (2021). Understanding Employee Responses to Software Robots: A Systematic Literature Review, p. 1358. Austin, TX, USA: Association for Information Systems (AIS).

Shrestha, Y. R., S. M. Ben-Menahem and G. von Krogh. (2019). "Organizational Decision-Making Structures in the Age of Artificial Intelligence." *California Management Review 61* (4), 66–83.

Sturm, T., J. Gerlacha, L. Pumplun, N. Mesbah, F. Peters, C. Tauchert, P. Buxmann. (2021). "Coordinating Human and Machine Learning for Effective Organization Learning." *MIS Quarterly 45* (3), 1581–1602.

Sturm, T. and F. Peters. (2020). The Impact of Artificial Intelligence on Individual Performance: Exploring the Fit between Task, Data, and Technology. *Proceedings of the 28th European Conference on Information Systems (ECIS 2020).*

Swets, J. and D. Green. (1963). *Signal Detection By Human Observers.*, First Edit, Vol. 19. Peninsula Publishing.

Teodorescu, M., L. Morse, Y. Awwad and G. Kane. (2021). "Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation." *MIS Quarterly 45* (3), 1483–1500.

Verghese, P. (2001). "Visual search and attention: A signal detection theory approach." *Neuron 31* (4), 523–535.

Wang, Y., L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, H. Liu. (2018). "Clinical information extraction applications: A literature review." *Journal of Biomedical Informatics 77,* 34–49.

Wang, Y., M. Xiong and H. G. T. Olya. (2020). Toward an understanding of responsible artificial intelligence practices. In: *Proceedings of the Annual Hawaii International Conference on System Sciences*, Vol. 2020-Janua, pp. 4962–4971.

Wolfe, J. M. and T. S. Horowitz. (2004). “What attributes guide the deployment of visual attention and how do they do it?” *Nature Reviews Neuroscience 5* (6), 495–501.

Yang, Q., A. Steinfeld, C. Rosé and J. Zimmerman. (2020). Re-examining Whether, Why, and How Human-AI Interaction Is Uniquely Difficult to Design. In: *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–13. New York, NY, USA: ACM.

Zhang, P., I. Benbasat, J. Carey, F. Davis, D. F. Galletta and D. Strong. (2002). “Human-Computer Interaction Research in the MIS Discipline.” *Communications of the Association for Information Systems 9* (20), 334–355.

Zhang, P. and N. Li. (2004). “An assessment of human-computer interaction research in management information systems: Topics and methods.” *Computers in Human Behavior 20* (2), 125–147.